Apuntes de Econometría

BORRADOR: No Citar-No Reproducir

Francisco A. Ramírez de León

2022-08-31

Advertencias:

1. Estas notas no sustituyen las lecturas asignadas en la clase de los libros de referencia.
2. Estarán siendo continuamente revisadas y se agradece cualquier observación de errores y mejoras explicativas de las demostraciones.
3. Cualquier error es de mi responsabilidad.

I. Introducción:

1.1. ¿De qué trata la econometría?

La econometría comprende un conjunto de herramientas estadísticas y conceptuales utilizadas en la evaluación empírica de las hipótesis provenientes de la teoría económica o en la estimación del impacto de la implementación de una política pública o privada. La existencia de este instrumental se debe a la necesidad de dotar de validación empírica a las relaciones causales que se deducen de los modelos económicos formulados con la teoría o a constatar si determinada política tiene o no el efecto esperado. En ese sentido, la econometría forma parte de la "caja de herramientas" del economista.

En términos formales, la econometría suele presentarse como la combinación de las herramientas estadísticas y matemáticas, con las hipótesis provenientes de la teoría económica.

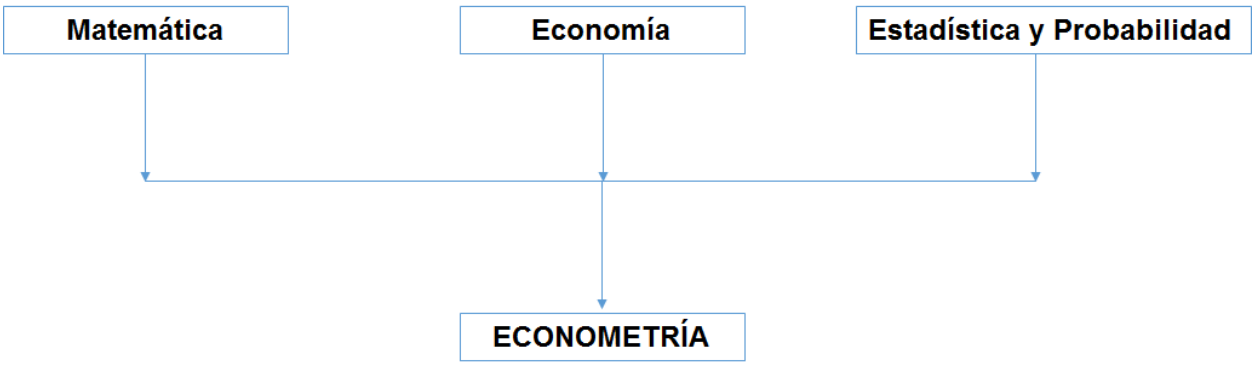


Figure 1: Econometría: Una combinación de conocimientos

El presente documento tiene como objetivo introducir a los estudiantes de economía al mundo de los métodos empíricos utilizados en el proceso de investigación en economía. La estrategia didáctica consiste en presentar ejemplos y cuestionar sobre los elementos a tomar en cuenta para agotar de manera exitosa la etapa empírica de un proyecto de investigación. Esto requiere: a) establecer la relación causal de interés b) seleccionar el

modelo apropiado, c) identificar la estrategia de estimacion apropiada al modelo seleccionado y d) seleccionar la estrategia de contraste adecuada y e) conocer el alcance y las limitaciones de los resultados obtenidos.

Todo comienza con una teoría (ya sea contable, económica o sociológica, por ejemplo). En economía, las teorías suelen estar expresadas en términos de funciones matemáticas, y representan la relación causal de interés. Por ejemplo:

$$Q = f(p, I)$$

Donde f es una función real, Q es la cantidad demandada de un bien, p el precio e I el ingreso.

En un curso de microeconomía se analiza el tipo de relación (positiva, negativa o nula) existente entre estas variables desde una perspectiva cualitativa. En la práctica, estamos interesados en saber "cuánto" se relacionan estas variables. Por lo que una definición de econometría es:

el uso de la teoría y datos de la economía, negocios, y de las ciencias sociales, junto con las herramientas de la estadística, para responder preguntas del tipo "cuánto"

Otros ejemplos de preguntas que expresan relación causal entre variables:

- ¿Cuánto incrementa el desempleo si el salario mínimo es incrementado en x%?
- ¿Cuánto sería el incremento de la inflación si el Banco Central decidiera incrementar la cantidad de dinero en 10%?
- ¿En cuánto se incrementarían las ventas de cemento si el precio fuese reducido en 3% y el gasto en publicidad aumentara en 2%?
- ¿Cuánto se reduciría la tasa de inasistencia escolar a nivel de primaria si se implementaran cuatro campañas de desparasitación en las escuelas públicas de X región?
- ¿Cuánto se reducirían los crímenes de calle (asaltos) si se duplicara el salario a los policías?

Entre otras preguntas muy interesantes.

1.2. Pasando del modelo económico al modelo econométrico:

En esta sección se sintetiza el proceso de especificación del modelo econométrico, con el propósito de mostrar las etapas involucradas en el mismo.

Son múltiples los factores que pueden incidir en una variable sobre la cuál se tenga interés. Como recuerdan, los modelos económicos son una simplificación de la realidad. La teoría económica describe el comportamiento 'promedio' o 'sistemático' del agente de interés.

Una forma de pensar el paso del modelos económico al modelo econométrico es proponiendo que el comportamiento observado en los datos es la suma de ese componente sistemático y un componente impredecible y aleatorio. Es decir,

$$C = f(p, I) + e$$

¿Qué representa e?

- Todos los factores omitidos del modelo simple
- Incertidumbre

Para terminar de obtener (*especificar*) el modelo econométrico, hay que proponer una forma funcional para f. La propuesta mas común es una función lineal:

$$f(p, I) = \beta_0 + \beta_1 p + \beta_2 I$$

Por lo que el modelo econométrico que vamos a estudiar es del tipo

$$C = \beta_0 + \beta_1 p + \beta_2 I + e$$

Donde β_0, β_1 y β_2 son parámetros desconocidos que estimaremos usando datos económicos y una técnica econométrico.

La forma funcional representa una hipótesis acerca de la relación entre las variables y hay que determinar si es compatible con la teoría económica y los datos. Con el uso del modelo econométrico y los datos, se hace inferencias acerca del mundo real y se aprende acerca del mismo en este proceso.

A este proceso de aprendizaje usando los elementos mencionados se le denomina Inferencia Estadística e incluye:

- Estimación de los parámetros económicos
- Predicción de los resultados económicos
- Contraste (testing) de hipótesis económicas.

1.3. Importancia:

Si la economía es considerada una ciencia (social) es debido a que los fenómenos que estudia son abordados desde la perspectiva del método científico, el cual considera como parte fundamental el contraste con la evidencia empírica de las hipótesis sobre las relaciones causales que caracterizan el fenómeno de interés, ya sea directamente a través de un experimento que reproduzca el fenómeno o indirectamente a través de su análisis a partir de información compilada relativa al fenómeno. La econometría es el conjunto de métodos y herramientras para realizar de manera sistemática dicho contraste y forma parte fundamental de la formación de un economista. Es la que en parte conecta lo que se aprende en los cursos de economía con lo que se aplica en la práctica.

1.4. Alcance:

No se limita a economía, sino a otras áreas: análisis de negocios, mercadeo, finanzas, así como al resto de las ciencias sociales.

2. Estructura de los datos económicos

2.1. Fuentes de los datos económicos:

Los datos provienen de dos fuentes:

1. Datos experimentales

Provienen de experimentos controlados. Escasos en economía. El valor de las variables independientes es fijo en muestras repetidas, permitiendo la repetición del fenómeno de interés bajo las mismas condiciones.

2. Datos no experimentales u observacionales

Provienen de encuestas. Datos de todas las variables son colectadas simultáneamente.

2.2. Características de los datos económicos:

Dependiendo del nivel de agregación:

- Micro-datos: individuos, hogares y firmas.
- Macro-datos: agregación de individuos, hogares y firmas a nivel local o nacional.

Pueden representar un acervo o un flujo:

- Flujo: resultado medido en un periodo de tiempo.
- Acervo: medido en un punto particular del tiempo.

Puede ser cuantitativa o cualitativa

- Cuantitativa: en términos numéricos.
- Cualitativos: en términos de clasificación o características.

2.3. Tipos de datos en economía:

Series de tiempo: datos recolectados en un intervalo discreto de tiempo.

El intervalo de tiempo varía acorde con la recurrencia en que se registran los datos. A esta variación se le denomina frecuencia. Existen los datos de "alta frecuencia o frecuencia alta" como datos semanales, diarios, por hora, y así sucesivamente. Asimismo, los datos de "baja frecuencia o frecuencia baja", tales como trimestrales, semestrales, anuales, bianuales y así sucesivamente.

La figura 2.2 muestra ejempl(os para series económicas de distinta frecuencia para el caso de la República Dominicana.

Como veremos en la sección de modelos para datos de series de tiempo, el comportamiento de este tipo de variables varía acorde con sus características: tendencia, estacionalidad, entre otras.

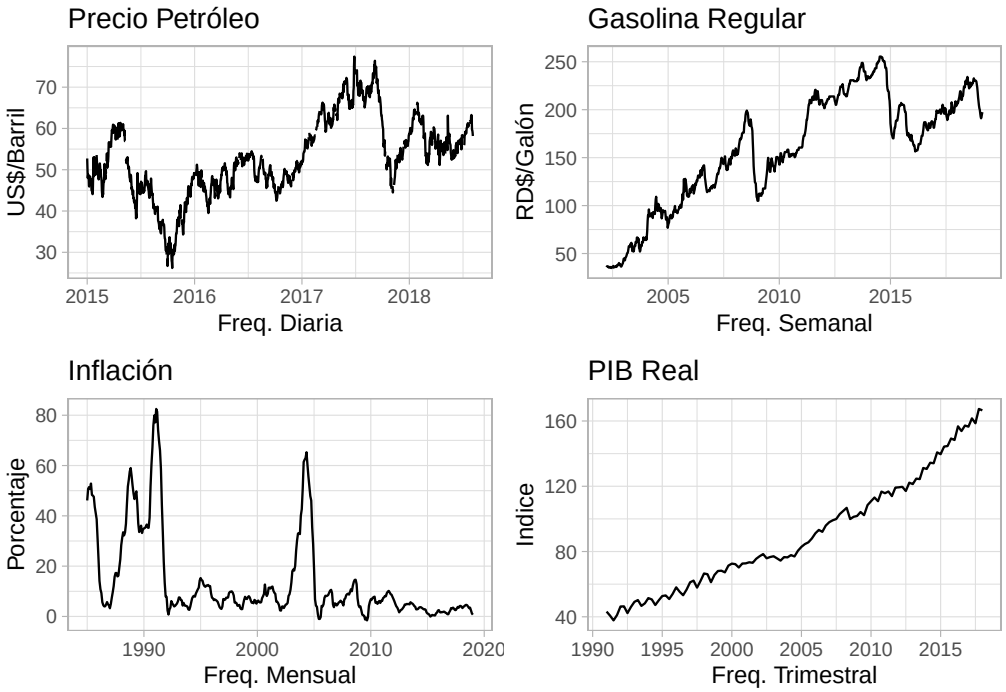


Figure 2: Ejemplo de datos de series de tiempo para distintas frecuencias

Corte transversal: datos recolectados a partir de un número determinado de unidades en un periodo de tiempo.

id	Salario	Educación	Experiencia
1	260.0	9	3
2	475.0	11	30
3	285.0	12	6
4	472.0	10	31
6	998.0	12	26
7	461.0	12	15
8	562.0	10	23
9	700.0	16	3
10	600.0	16	3
11	779.0	12	24
12	700.0	12	21
13	998.0	12	26
14	900.0	17	15
...	...	...	...

Figure 3: Ejemplo de corte transversal

Panel o longitudinales: datos recolectados a partir de un número de unidades en intervalos discretos de tiempo.

id	Periodo	Salario	Educación	Experiencia
1	1	260.0	9	3
1	2	305.0	9	4
1	3	402.0	9	5
2	1	475.0	11	30
2	2	500.0	11	31
2	3	525.0	11	32
3	1	285.0	12	6
3	2	624.0	12	7
3	3	698.0	12	8
4	1	472.0	10	31
4	2	512.0	10	32
4	3	545.0	10	33
...	...	...	...	...

Figure 4: Ejemplo de datos de panel

El Modelo de Regresión Lineal.

El análisis empírico de la relación causal propuesta por la teoría es realizado a partir de la especificación del modelo de regresión lineal (MRL). Los supuestos en los que se basa el MRL permiten la obtención de la denominada función de esperanza condicional, que especifica el comportamiento del valor promedio o esperado de nuestra variable de interés en función de nuestra variable causal y nuestros controles. Los supuestos son los siguientes:

Supuesto 1: Cada observación y_i es explicada por el componente sistemático $E(y_i|x_{i1},...,x_{ik},\beta)$ y el componente sistemático ε_i:

$$y_i = E(y_i|x_{i1},...,x_{ik},\beta) + \varepsilon_i$$

donde $E(y_i|x_{i1},...,x_{ik},\beta)$ se conoce como la *función de regresión* o *esperanza condicional* , la cual se asume que es lineal:

$E(y_i|x_{i1},...,x_{ik},\beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$

Por esta razón, este supuesto suele estar presentado como: *la forma funcional es lineal en los parámetros y el término error entra de manera aditiva, por lo que cada observación viene descrita por:*

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i$$

Ejemplo ¿Existe relación entre acceso a un seguro de salud y las condiciones de salud?

En este caso, la pregunta de investigación establece una relación causal del tipo $seguro \rightarrow salud$, donde *seguro* es una variable explicativa relativa a la información de tenencia de algun tipo de afiliación de seguro de salud y *salud* algún indicador de bienestar del individuo.

De acuerdo al supuesto 1, la relación causal puede ser especificada como:

$salud_i = \beta_0 + \beta_1 seguro_i + \varepsilon_i$

Supuesto 2: Supuesto de Identificación.

Para cada observación i,

$$E[\varepsilon_i|x_{i1},...,x_{ik}] = 0$$

Este supuesto permite la identificación del efecto causal, es decir, garantiza la exogeneidad de las variables explicativas respecto al error del modelo. Es decir, tomando derivada parcial respecto a la variable x_{ij}

$$\frac{\partial E[y_i|x_{i1}],...,x_{ik}}{\partial x_{ij}} = \beta_j + \frac{\partial E[\varepsilon_i|x_{ij},...,x_{ik}]}{\partial x_{ij}} = \beta_j + 0 = \beta_j$$

Es decir, la información contenida en el error del modelo no esta correlacionada con x_{ij}. En ese sentido es que se dice que esta es exógena (respecto al ε_i). Sin este supuesto note que no es posible identificar el efecto causal, puesto que el segundo término no sería igual a cero. Un caso se muestra en el siguiente ejemplo.

Ejemplo ¿Existe relación entre acceso a un seguro de salud y las condiciones de salud?

En nuestro ejemplo de la relación entre seguro y condiciones de salud, suponga que un tercer factor no observable tiene un efecto causal sobre la salud:

$$factorX \rightarrow salud$$

y que a la vez tiene se correlaciona con la demanda de seguros

$$factorX \rightarrow seguro$$

Como esta variable ($factorX$) no es observable, está implícitamente considerada en ε, por lo tanto, se deduce que en este caso

$$E[\varepsilon_i | x_{i1}, ..., x_{ik}] \neq 0$$

Es decir,

$$\frac{\partial E[salud_i | seguro_i}{\partial seguro_i} = \beta_1 + \frac{\partial E[\varepsilon_i | seguro_i]}{\partial seguro_i} \neq \beta_1$$

No se cumple el supuesto 2.

Usualmente, los economistas quieren más que solo encontrar la mejor aproximación lineal de una variable, dado el conjunto de regresores. Ellos quieren relacionaes económicas que son generalmente más válidas que la muestra de donde provengan. Ellos quieren extraer conclusiones acerca de qué pasaría si una de las variables cambia. Esto es, ellos quieren decir algo acerca de los valores que no están incluidos en la muestra. Por ejemplo, se quisiera predecir el salario de un individuo sobre la base de sus antecedendes y características y determinar cómo sería diferente si esta persona tiene más años de educación. En este caso, se quiere que la relación encontrada sea más que una concidencia histórica: es decir, refleje una relación fundamental. Para hacer esto se asume que hay una relación general que es válida para todas las posibles observaciones de una población bien definida (ejemplo, todos los individuos con un trabajo pagado en una fecha determinada, o todas las empresas de una industria.). Restringiendo la atención en relacionaes lineales, se especifica un **modelo estadístico** como:

$$y_i = \beta_1 + \beta_2 x_2 + ... + \beta_K x_{iK} + \epsilon_i \tag{3.34}$$

o

$$y_i = x_i' \beta + \epsilon_i \tag{3.35}$$

donde y_i y x_i son variables observables y ϵ_i es no observable y se refiere al **término error** o término de perturbación. En este contexto, y_i se refiere a la variable dependiente y las variables en x_i son llamadas variables independientes, variables explicativas, regresores u, ocasionalmente, covariantes. Los elementos en β son los parámetros poblacionales desconocidos. La igualdad en (3.35) se supone que se cumple para cada una de las posibles observaciones, mientres nosotros solo observamos una **muestra** de N observaciones. Se considera esta muestra como una realización de todas las muestras potenciales de tamaño N que pueden ser extraídas de la misma población. En este sentido, y_i y ϵ_i (y con frecuencia x_i) se consideran como **variables aleatorias**. Cada observación corresponde a una realización de estas variables aleatorias. De nuevo, se puede usar notación matricial y crear un vector con todas las observaciones para escribir:

$$y = X\beta + \epsilon \tag{3.36}$$

donde y y ϵ son vectores N-dimensionales, y X, como antes es de dimensióm $N \times K$. Note la diferencia entre esta ecuación y la (3.29).

En contraste con (3.15) y (3.29), (3.35) y (3.36) son relaciones poblacionales, donde β es un vector de parámetros desconocidos que caracterizan la población. El **proceso de muestreo** describe como la muestra fue tomada. En una primera aproximación, las x_i son consideradas como fijas y no estocásticas, que significa que cada nueva muestra tiene la misma matriz X. En este caso se dice que las x_i son **determinísticas**. Una nueva muestra solo implica nuevos valores para ϵ_i, o de manera equivalente para y_i. El caso donde esta condición es interesante es en un laboratorio, donde un investigador puede fijar las condiciones de un experimento específico (por ejemplo, temperatura, presión del aire). En economía se trabaja típicamente con datos no experimentales. A pesar de esto, es conveniente y en casos particulares apropiada en un contexto económico actual como si las x_i son determinísticas. En este caso, se tienen que hacer supuestos sobre la distribución muestral de ϵ_i . Uno conveniente corresponde a **muestreo aleatorio**, donde cada error ϵ_i es una extracción aleatoria de una distribución de la población, independiente de los otros términos de error.

En una segunda mirada, una nueva muestra implica nuevos valores para ambos x_i y ϵ_i, de tal manera que en cada momento un nuevo conjunto de N para (y_i, x_i) es extraído. En este caso muestreo aleatorio significa que cada conjunto (y_i,x_i) es una extracción aleatoria de la distribución de la población. En este contexto, resulta importante hacer supuestos acerca de la distribución conjunta de x_i y ϵ_i, en particular relativo al grado en que la distribución de ϵ_i depende de X. La idea de muestra aleatoria es más entendible en el contexto de corte transversal, donde el interés recae en una población grande y fija. Por ejemplo, todos los hogares dominicanos en Enero de 2015. En el contexto de series de tiempo, diferentes observaciones se refieren a diferentes periodos, y no hace sentido tener una muestra aleatoria de periodos. En su lugar, se puede tomar una visión que la muestra que se tiene es una realización de lo que hubiese pasado en un momento dado y la aleatoriedad se refiere a estados alternativos del mundo. En ese caso se puede hacer supuestos acerca de la forma en que los datos han sido generados (en lugar de la forma en que los datos han sido muestreados).

Es importante darse cuenta que sin supuestos adicionales el modelo estadístico (3.35) es una tautología: para cualquier valor de β se puede definir un conjunto de ϵ_i tal que (3.35) se cumpla para cada observación. Se necesita imponer algunos supuestos para darle al modelo un significado. Un supuesto común es que el valor esperado de ϵ_i dadas todas las variables explicativas en x_i es cero, esto es:

$$E\{\epsilon_i | x_i\} = 0 \tag{3.37}$$

Usualmente, las personas se refieren a este supuesto diciendo que las variables son **exógenas**. Bajo este supuesto:

$$E\{y_i | x_i\} = x_i\beta \tag{3.38}$$

de tal manera que la línea de regresión (poblacional) $x_i\beta$ descibe la expectativa condicional de y_i dado los valores para x_i. Los coeficientes β_k mide como el valor esperado de y_i es afectado si el valor de x_{ik} cambia, manteniendo los demas elementos en x_i constantes (la condición de **ceteris paribus**). La teoría económica, sin embargo, con frecuencia sugiere que el modelo (3.35) describe una relación causal, en la cual los coeficientes β miden los cambios en y_i causados por un cambio ceteris paribus en x_{ik}. En esos casos, ϵ_i tiene una interpretación económica (no solamente estadística) e imponer que esta no está correlacionada con x_i, como se impone $E\{\epsilon_i | x_i\} = 0$, puede no estar justificado. Debido que en muchas aplicaciones se puede argumentar que las variables no observables en el término error están relacionadas con las observables en x_i, se debe ser cauteloso interpretando los coeficientes de regresión como midiendo efectos causales.

Ahora que los coeficientes β tienen significado, se puede tratar de usar la muestra (y_i, x_i), $i = 1, ..., N$, para decir algo acerca de ellos. La regla que dice como una muestra dada es traducida en un valor apropiado para β es lo que se llama un **estimador**. El resultado para un muestra es llamado un **estimado**. El estimador

es un vector de variables aleatorias, debido a que la muestra puede cambiar. El estimado es un vector de números. El estimador más utilizado en econometría es el estimador de **mínimos cuadrados ordinarios**. Este es la regla de mínimos cuadrados descrito en la sección anterior aplicado a una muestra disponible. El estimador de MCO para β esta dado por:

$$b = \Big(\sum_{i=1}^{N} x_i x_i'\Big)^{-1} \sum_{i=1}^{N} x_i y_i \tag{3.39}$$

Dado que hemos asumido un modelo subyacente "verdadero" (3.35), combinado con una muestra, b es ahora un vector de variables aleatorias. Nuestro interés recae en el vector β de parámetros verdaderos pero desconocidos, y b es considerado una aproximación a este. Mientras una muestra dada solo produce un estimador puntual, se evalua la calidad de este a través de las propiedades subyacentes del estimador. El estimador b tiene una distribución muestral debido a que su valor depende de la muestra que ha sido extraída aleatoriamente de la población.

Es extremadamente importante entender la diferencia entre el estimador b y los valores poblacionales β. El primero es un vector de variables aleatorias, el resultado depende de la muestra que se empleó. El segundo es un conjunto de números fijos desconocidos, que caracteriza el modelo poblacional. Asimismo, la distinción entre el término error ϵ_i y los residuales e_i es importante. Los términos de error son no observables, y supuestos sobre su distribución acerca de estos son necesarios para derivar las propiedades muestrales de los estimados de β. Los residuales se obtienen despues de la estimación, y los valores dependen del estimado de β y en consecuencia depende de la muestra y el método de estimación. Las propiedades del término error ϵ_i y los residuos e_i no son las mismas y ocasionalmente son muy distintas.

El modelo de regresión lineal en combinación con el método de mínimos cuadrados ordinarios(MCO) es uno de los pilares de la econometría. En la primera parte de estos apuntes se revisa el modelo de regresión lineal y sus supuestos, cómo este puede ser estimado, evaluado e interpretado y cómo puede ser utilizado para generar predicciones para contrastar hipótesis económicas.

El capítulo empieza introduciendo el método de mínimos cuadrados ordinarios como herramienta algebraica, en lugar de estadística. Esto es debido a que MCO tiene la propiedad atractiva de proveer la mejor aproximación lineal, independientemente de como los datos han sido generados, o los supuestos considerados. El Modelo de Regresión Lineal (MRL) es introducido en la sección XX mientras la sección XX discute las propiedades del estimador de MCO en este modelo bajo los supuestos de Gauss-Markov (GM). La sección XX discute medidas de bondad de ajuste para el modelo lineal, y el contraste de hipótesis es tratado en la sección XX. En la sección XX nos movemos a los casos donde los supuestos de GM no necesariamente se satisfacen y las propiedades de muestra pequeña del estimador de MCO son desconocidas. En dichos casos, el comportamiento limitado del estimador de MCO cuando - hipotéticamente- el tamaño de muestra se vuelve infinitamente grande comúnmente utilizadas para aproximar las propiedades de muestra pequeña. La sección XX provee un ejemplo. En las secciones XX y XX se discuten los problemas de datos relativos a la multicolinealidad, valores atípicos y valores no observados, mientras que en la sección XX se da atención a la predicción usando el modelo de regresión lineal. Se proveen ejemplos para interpretar los coeficientes del MRL, como contrastar los supuestos del modelo etc.

Algebra de Mínimos Cuadrados

Mínimos cuadrados ordinarios

Suponga que tiene una muestra con N observaciones de salarios de individuos y un número de características de estos individuos, tales como género, años de educación y experiencia. El interés principal recae en la pregunta de cómo en esta muestra los salarios están relacionados a las otras variables observables. Denote los salarios por y (el regresando) y las demás K-1 características por $x_2, ..., x_K$ (los regresores). Más abajo se aclara por qué esta numeración de las variables es conveniente. Ahora se hace la pregunta: ¿cuál es la combinación lineal de $x_2, ...x_N$ y una constante que da una buena aproximación de y? Para responder esta

pregunta, primero considere una combinación lineal arbitraria, incluyendo una constante, que puede ser escrita como:

$$\hat{\beta}_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_K x_K \tag{3.1}$$

donde $\hat{\beta}_1, ..., \hat{\beta}_K$ son constante por ser seleccionadas. Ahora indexe las observaciones por i, tal que $i = 1, \cdots, N$. Ahora la diferencia entre los valores observado de y_i y su aproximación lineal es

$$y_i - [\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_K x_{iK}] \tag{3.2}$$

Para simplificar las derivaciones se introduce una notación mas resumida. En el Capítulo 1 se proveen detalles para los lectores no familiarizados con notación de vector. El caso especial es cuando $K = 2$ que discutiremos en la siguiente sección. Para un K general se coleccionan los valores x para los individuos i en un vector x_i, que incluye una constante. Esto es,

$$x_i = \begin{pmatrix} 1 & x_{i2} & x_{i3} & \cdots & x_{iK} \end{pmatrix}' \tag{3.3}$$

donde ' es usado para denotar la transpuesta. Coleccionado los coeficientes $\hat{\beta}$ en un vector K dimensional $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_K \end{pmatrix}'$ se puede brevemente reescribir como:

$$y_i - x_i'\hat{\beta} \tag{3.4}$$

Se desea elegir los valores para $\hat{\beta}_1, \cdots, \hat{\beta}_K$ tal que las diferencias son pequeñas. A pesar de que distintas medidas se pueden usar para definir por lo que significa "pequeñas", la más común es elegir los $\hat{\beta}$ tal que la suma de las diferencias al cuadrado es lo mas pequeña posible. En este caso, se determina $\hat{\beta}$ tal que minimice la siguiente función objetivo:

$$S(\hat{\beta}) \equiv \sum_{i=1}^{N} (y_i - x_i'\hat{\beta})^2 \tag{3.5}$$

Esto es, se minimiza la suma de los errores al cuadrado aproximados. Este enfoque es definido como mínimos cuadrados ordinarios. Tomando cuadrados se asegura de que desviaciones positivas y negativas no se cancelen entre sí cuando se tome la sumatoria.

Para resolver el problema de minimización, considere las condiciones de primer orden, obtenidas diferenciando $S(\hat{\beta})$ respecto al vector $\hat{\beta}$.

$$\min S(\hat{\beta}) \equiv \sum_{i=1}^{N} (y_i - x_i'\hat{\beta})^2 \tag{3.6}$$

Esto da las siguientes K condiciones

$$-2 \sum_{i=1}^{N} x_i (y_i - x_i'\hat{\beta}) = 0 \tag{3.7}$$

Dividiendo de ambos lados por -2:

$$\sum_{i=1}^{N} x_i (y_i - x_i'\hat{\beta}) = 0 \tag{3.8}$$

Resolviendo la sumatoria:

$$\sum_{i=1}^{N} x_i y_i - \Big(\sum_{i=1}^{N} x_i x_i' \Big) \hat{\beta} = 0 \tag{3.9}$$

Despejando:

$$\left(\sum_{i=1}^{N} x_i x_i' \right) \hat{\beta} = \sum_{i=1}^{N} x_i y_i \tag{3.10}$$

donde:

$$\left(\sum_{i=1}^{N} x_i x_i' \right) = \sum_{i=1}^{N} \left\{ \begin{pmatrix} 1 \\ x_{i2} \\ x_{i3} \\ \cdots \\ x_{iK} \end{pmatrix} \begin{pmatrix} 1 & x_{i2} & x_{i3} & \cdots & x_{iK} \end{pmatrix} \right\} \tag{3.11}$$

$$= \sum_{i=1}^{N} \left\{ \begin{pmatrix} 1 & x_{i2} & x_{i3} & \cdots & x_{iK} \\ x_{i2} & x_{i2}^2 & x_{i2}x_{i3} & \cdots & x_{i2}x_{iK} \\ x_{i3} & x_{i3}x_{i2} & x_{i3}^2 & \cdots & x_{i3}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{iK} & x_{iK}x_{i2} & x_{iK}x_{i3} & \cdots & x_{iK}^2 \end{pmatrix} \right\}$$

$$= \begin{pmatrix} N & \sum_{i=1}^{N} x_{i2} & \sum_{i=1}^{N} x_{i3} & \cdots & \sum_{i=1}^{N} x_{iK} \\ \sum_{i=1}^{N} x_{i2} & \sum_{i=1}^{N} x_{i2}^2 & \sum_{i=1}^{N} x_{i2}x_{i3} & \cdots & \sum_{i=1}^{N} x_{i2}x_{iK} \\ \sum_{i=1}^{N} x_{i3} & \sum_{i=1}^{N} x_{i3}x_{i2} & \sum_{i=1}^{N} x_{i3}^2 & \cdots & \sum_{i=1}^{N} x_{i3}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{iK} & \sum_{i=1}^{N} x_{iK}x_{i2} & \sum_{i=1}^{N} x_{iK}x_{i3} & \cdots & \sum_{i=1}^{N} x_{iK}^2 \end{pmatrix}$$

Asimismo,

$$\sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} \left\{ \begin{pmatrix} 1 \\ x_{i2} \\ x_{i3} \\ \cdots \\ x_{iK} \end{pmatrix} y_i \right\} = \sum_{i=1}^{N} \begin{pmatrix} y_i \\ y_i x_{i2} \\ y_i x_{i3} \\ \cdots \\ y_i x_{iK} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i x_{i2} \\ \sum_{i=1}^{N} y_i x_{i3} \\ \cdots \\ \sum_{i=1}^{N} y_i x_{iK} \end{pmatrix}$$

Estas ecuaciones son a veces llamadas **ecuaciones normales**. Como este sistema tienen K incógnitas (el número de parámetros desconocidos), se puede obtener una solución única para $\hat{\beta}$ provisto que la matriz simétrica $\sum_{i=1}^{N} x_i x_i'$, que contiene la suma de las sumas de cuadrados y productos cruzados de los regresores x_i pueda ser invertible. Por el momento, se asume que es el caso. La solución del problema de minimización, que puede ser denotada por b, esta dada por:

$$b = \left(\sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i y_i \tag{3.12}$$

Revisando las condiciones de segundo orden, se verifica que b corresponde a un mínimo. Derivando las condiciones de primero orden respecto a $\hat{\beta}$:

$$\frac{\partial^2 S}{\partial \hat{\beta}^2} = \Big(\sum_{i=1}^{N} x_i x_i' \Big) \tag{3.13}$$

que es una matriz positiva-definida.

La combinación lineal resultante de x_i esta dada por:

$$\hat{y}_i = x_i' b \tag{3.14}$$

que es la **mejor aproximación lineal** de y con $x_2, \cdots, x_K$ y una constante. La frase "mejor" refiere al hecho de que la suma de diferencias al cuadrado entre los valores observados y_i y valores ajustados $\hat{y}_i$ es mínima para solución de mínimos cuadrados b.

Derivando la aproximación lineal, no se ha usado ninguna teoría económica o estadística. Es simplemente una herramienta algebraica, y se cumple irrespectivamente la forma en que los datos fueron generados. Esto es, dado un conjunto de variables se puede determinar la mejor aproximación lineal de una variable usando las otras variables. El único supuesto que se hizo (que es directamente verificable en los datos) es que la matriz $KxK \sum_{i=1}^{N} x_i x_i'$ es invertible.

Esto significa que ninguna de las x_{ik} es una combinación lineal exacta de las otras y por tanto redundante. Esto es usualmente como el supuesto de **no-multicolinealidad**. Puede estresarse que la aproximación lineal es un resultado dentro de muestra (esto es, en principio no da información acerca de las observaciones (individuos) que no están incluidos en la muestra), y en general, no hay una interpretación directa de los coeficientes.

A pesar de estas limitaciones, los resultados algebraicos sobre e método de mínimos cuadrados son muy útiles. Definiendo un **residuo** e_i como la diferencia entre el valor observado y el aproximado, $e_i = y_i - \hat{y}_i = y_i - x_i' b$ se puede descomponer la observable y_i como:

$$y_i = \hat{y}_i + e_i = x_i' b + e_i \tag{3.15}$$

Esto permite escribir el valor mínimo para la función objetivo como:

$$S(b) = \sum_{i=1}^{N} e_i^2 \tag{3.16}$$

la cual es conocida como la **suma de residuos al cuadrado**. Se puede mostrar que el valor aproximado $x_i' b$ y el residuo e_i satisfacen ciertas propiedades por construcción. Por ejemplo, si se reescribe (3.8), sustituyendo la solución de MCO para b, se obtiene

$$\sum_{i=1}^{N} x_i (y_i - x_i b) = \sum_{i=1}^{N} x_i e_i = 0 \tag{3.17}$$

Esto significa que el vector $e = (e_1, ..., e_N)$ es ortogonal a cada vector de observaciones de la variable x. Por ejemplo, si x_i contiene una constante, implica que $\sum_{i=1}^{N} e_i = 0$. Esto es, el promedio del residuo es cero. Este es un resultado atractivo. Si el promedio de los residuos no fuera cero, esto significa que se puede mejorar la aproximación añadiendo o substrayendo la misma constante para cada observación, esto es, cambiando b_1. En consecuencia, para la observación promedio se tiene

$$\bar{y} = \bar{x}' b \tag{3.18}$$

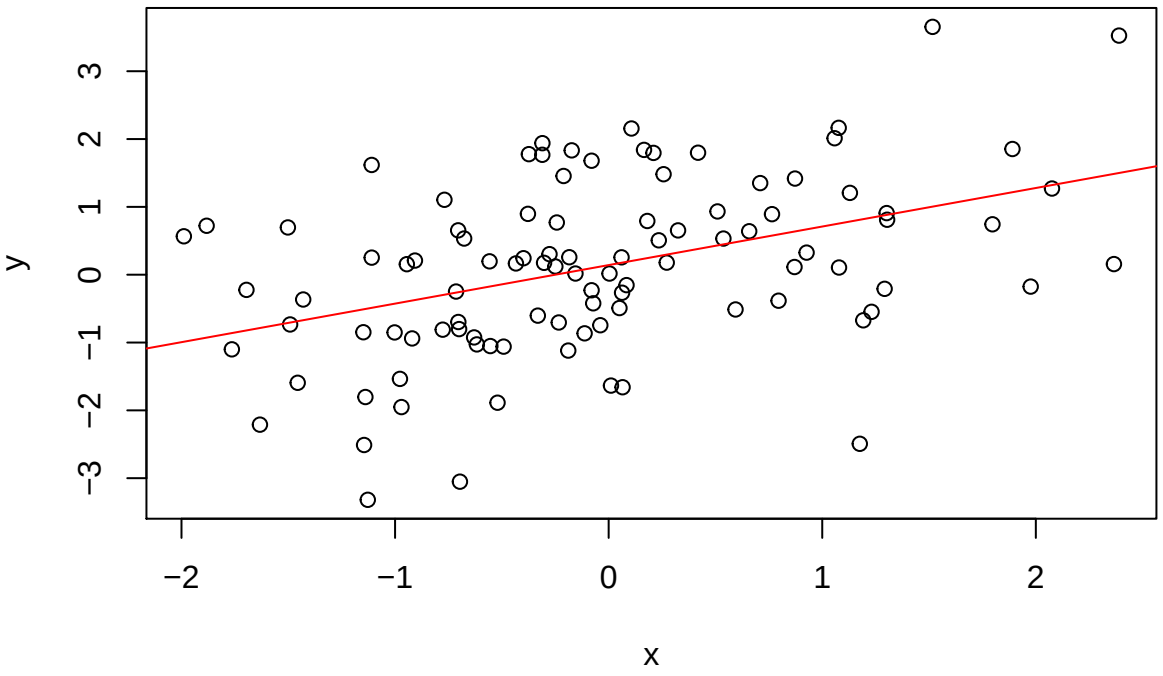
donde $\bar{y} = (1/N)\sum_{i=1}^{N} y_i$ y $\bar{x} = (1/N)\sum_{i=1}^{N} x_i$ un vector K dimensional de medidas muestrales. Esto muestra que para la observación promedio hay un error de aproximación. Interpretación similar existe para otros regresores: si la derivada de la suma al cuadrado de los errores aproximados respecto a $\hat{\beta}$ es positivo, esto es si $\sum_{i=1}^{N} x_{ik} e_i > 0$, significa que se puede mejorar la función objetivo bajando a $\hat{\beta}$. La ecuación (3.15) descompone los valores observados de y_i entre dos componentes ortogonales: el valor ajustado (relacionado a x_i) y el residuo.

Regresión Lineal Simple

En el caso de $K = 2$ solo tenemos un regresor y una constante. En este caso las observaciones (y_i, x_i) se pueden mostrar en un gráfico bidimensional con los valores de x en el eje horizontal y los valores de y en el eje vertical. Esto es presentado en la Figura 1 para una base de datos hipotética

```
x <- rnorm(100)
e <- rnorm(100)
y <- 0.1 + 0.5*x+e
plot(x,y,main="Figura 1: Regresión Lineal Simple")
abline(lm(y~x),col="red")
```

Figura 1: Regresión Lineal Simple



La mejor aproximacion lineal de y con x y una constante es obtenida minimizando la suma de residuos al cuadrado, la cual - en este caso bidimensional - es igual a la distancia vertical entre una observacion y el valor ajustado. Todos los valores ajustradao estan sobre la linea recta, la **linea de regresión**.

Debido a que una matriz 2x2 puede ser invertida analíticamente, se puede derivar soluciones para b_1 y b_2 en este caso especial de la expresión general ya vista. Equivalentemente, podemos minizar la suma de residuos al cuadrado respecto a las incógnitas directamente. Entonces, se tiene:

$$S(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{N} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{3.19}$$

Las condiciones de primer orden son:

$$\frac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{N}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{3.20}$$

$$\frac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{N} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{3.21}$$

De las expresiones anteriores se obtiene

$$b_1 = \frac{1}{N}\sum_{i=1}^{N} y_i - b_2 \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{y} - b_2 \bar{x} \tag{3.22}$$

Resolviendo para b_2

$$\sum_{i=1}^{N} x_i y_i - b_1 \sum_{i=1}^{N} x_i - \left(\sum_{i=1}^{N} x_i^2\right) b_2 = 0 \tag{3.23}$$

Sustituyendo:

$$\sum_{i=1}^{N} x_i y_i - N\bar{x}y - \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right) b_2 = 0 \tag{3.24}$$

Despejando:

$$b_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3.25}$$

Dividiendo el numerador y el denominado por $N-1$ se tiene que la solucion de MCO b_2 es la razón de la covarianza muestral entre y y x y la varianza muestral de x. De (3.22), el intercepto es determinado tal que el promedio de errores de proximacion es igual a cero.

Notación Matricial

Debido a que los econometristas hacen uso frecuente de expresiones con matrices para notación resumida, alguna familiaridad con el lenguaje de matrices es un prerequisito para leer literatura econométrica. En estos apuntes, se rehacen los resultados usando notación matricial y ocasionalmente, cuando la alternativa es complicada, se restringe la atención a expresiones con matrices solamente. Usando matrices, derivando la solucion de mínimos cuadrados es mas rápida, pero requiere algún conocimiento de cálculo diferencial matricial. Se introduce la siguiente notación:

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_N' \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

En la matriz X, una matriz $N \times K$ la fila i se refiere a la observación i, y la columna k se refiere al regresor o variable explicativa k. El criterio a minimizar puede ser reescrito en notación matricial usando el hecho que el producto interno de un vector a con el mismo $(a'a)$ es la suma de sus elementos al cuadrado. Esto es:

$$S(\widetilde{\beta}) = (y - X\widetilde{\beta})'(y - X\widetilde{\beta}) = y'y - 2y'X\widetilde{\beta} + \widetilde{\beta}'X'X\widetilde{\beta} \tag{3.26}$$

de donde la solución se deriva diferenciando respecto a $\widetilde{\beta}$ e igualando a cero:

$$\frac{\partial S(\widetilde{\beta})}{\partial \widetilde{\beta}} = -2(X'y - X'X\widetilde{\beta}) = 0 \tag{3.27}$$

- Resolviendo se obtienen la solución de MCO:

$$b = (X'X)^{-1}X'y \tag{3.28}$$

que es exactamente igual a la definición en sumatorias, pero ahora en notación matricial.

Note que hay que asumir que $X'X = \sum_{i=1}^{N} x_i x_i'$ es invertible, esto es, que no hay multicolinealidad exacta (o perfecta).

Como antes, se puede descomponer y como:

$$y = Xb + e \tag{3.29}$$

donde e es un vector n-dimensional de residuos. Las condiciones de primer orden implican que

$$X'(y - Xb) = 0 \tag{3.30}$$

o

$$X'e = 0 \tag{3.31}$$

que significa que cada columna de la matriz X es ortogonal al vector de residuos.

De (3.28) se puede escribir (3.29) como:

$$y = Xb + e = X(X'X)^{-1}X'y + e = \hat{y} + e \tag{3.32}$$

Donde el valor predicho para y está dado por:

$$\hat{y} = Xb = X(X'X)^{-1}X'y = P_X y \tag{3.33}$$

En álgebra lineal, la matriz $P_X = X(X'X)^{-1}X'$ es conocida como una matriz de proyección. Proyecta el vector y sobre las columnas de X (espacio columna de X). Esta es justamente la translación geométrica de encontrar la mejor aproximación lineal de y a partir de las columnas (regresores) en X. La matriz P_X es tambien conocida como la "matriz sombrero" debido a que transforma y en $\hat{y}$. El vector de residuos de la proyección $e = y - Xb = (I - P_X)y = M_X$, es el complemento ortogonal. Es una proyección de y sobre el espacio ortogonal al espacio creado por las columnas de X. Esta intrepetación es usualmente útil. Por ejemplo, proyectando dos veces sobre el mismo espacio deja el resultado inalterado, tal que se cumple que $P_X P_X = P_X$ y $M_X M_X = M_X$. Más imortante, se tiene que $P_X M_X = 0$ debido a que el espacio columna de X y su complemento ortogonal no tienen algo en común (excepto el vector nulo). Esta es una forma alternativa para interpretar el resultado que $\hat{y}$ y e, y también X y e son ortogonales.

Propiedades de muestra pequeña

Los supuestos de Gauss-Markov

Si el estimador de MCO b provee una buena aproximación del vector de parámetros desconocidos β depende de manera crucial de los supuestos que se hacen sobre la distribución de ϵ_i y su relación con x_i. Un caso estándar en el que MCO tiene buenas propiedades es caracterizado por las condiciones de Gauss-Markov. Más adelante, en otros capítulos, se consideran condiciones más débiles bajo las cuales MCO aun tiene alguunas propiedades atractivas . Por ahora, es importante darse cuenta que las condiciones Gauss-Markov no son todas estrictamente necesarias para justificar el uso del estimador de MCO. Solo constituyen un caso simple en el cual las propiedades de muestra pequeña de b se derivan facílmente.

Dado el modelo de regresion lineal

$$y_i = x_i'\beta + \epsilon_i$$

Las condicones de Gauss - Markov son:

$$E\{\epsilon_i\} = 0 \tag{S1}$$

$$\{\epsilon_i, ..., \epsilon_N\} \;\; y \;\; \{x_1, ..., x_N\} \;\; son\ independientes \tag{S2}$$

$$V\{\epsilon_i\} = \sigma^2, \qquad i = 1, ..., N \tag{S3}$$

$$cov\{\epsilon_i, \epsilon_j\} = 0, \qquad i, j = 1, .., N, i \neq j. \tag{S4}$$

El supuesto $S1$ dice que el valor esperado del término error es cero, lo que significa que, en promedio, la linea de regresión es la correcta. El supuesto $S3$ establece que todos los términos error tienen la misma varianza, el cual es también llamado **homocedasticidad**, mientras que el supuesto $S4$ impone correlación cero entre los diferentes términos de error. Este excluye cualquier forma de **autocorrelación**.Juntos, los supuestos $S1$, $S3$ y $S4$ implican que el término error son extracciones no correlacionadas de una distribución con expectativa cero y varianza constante σ^2. Usando notación matricial, es posible reescribir estas tres condiciones como:

$$E\{\epsilon_i\} = 0$$
$$V\{\epsilon_i\} = \sigma^2 I_N$$

Donde I_N es la matriz identidad de dimensión $N \times N$. Esto dice que la matriz de covarianza del vector de errores es una matriz diagonal con σ^2 sobre la diagonal. El supuesto $S2$ implica que X y ϵ son independientes. Esto significa que conocer X no informa nada acerca de la distribución del término error. Este supuesto es fuerte e implica que:

$$E\{\epsilon_i|X\} = E\{\epsilon_i\} = 0 \tag{3.40}$$

y

$$V\{\epsilon_i|X\} = V\{\epsilon\} = \sigma^2 I_N \tag{3.41}$$

Esto es, la matriz de valores de los regresores X no provee ninguna información acerca de los valores esperados de los términos de error o su covarianza. Las condiones (3.39) y (3.40) combinan los elementos necesarios de los supuestos de Gauss-Markov necesarios para que los resultados que sigue sean válidos. Condicionando en X, se puede tomar a X como un vector no estocástico. La razón para esto, es que los resultados en la matriz X

puede ser tomado como dada sin afectar las propiedades de ϵ, esto es, se pueden derivar todas las proiedades condicional a X. Por simplicidad, se toma este enfoque. Bajo las condiciones de Gauss Markov $S1$ y $S2$, el modelo lineal puede ser interpreado como la expectativa condicional de y_i dado x_i, esto es, $E\{y_i|x_i\} = x_i'\beta$. Esta es una implicación directa de (3.39)

Propiedades del estimador de MCO

Bajo los supuestos $S1$-$S4$, el estimador de MCO de b para β tiene algunas propiedades deseables. Primero que todo, es **insesgado**. Esto significa que, en muestreo repetido, se puede espera que el estimador de MCO es en promedio igual al valor verdadero de β. Esto es: $E\{b\} = \beta$.

Demostración:

$$E\{b\} = \{(X'X)^{-1}X'y\} = E\{\beta + (X'X)^{-1}X'\epsilon\} \tag{3.42}$$

$$= \beta + E\{(X'X)^{-1}X'\epsilon\} = \beta \tag{3.43}$$

Este último paso es:

$$E\{(X'X)^{-1}X'\epsilon\} = E\{(X'X)^{-1}X'\}E\{\epsilon\} = 0 \tag{3.44}$$

debido al supuesto $S2$, X y ϵ son independientes y, de $S1$, $E\{\epsilon\} = 0$

Note que no se usaron los supuestos $S3$ y $S4$ en la demostración. Esto muestra que el estimador de MCO es insesgado en la medida que los términos de error tengan media cero y sean independientes de todas las variables explicativas, aun si están presentes la heteocedasticidad o la autocorrelación. Si un estimador es insesgado, esto siginifica que su distribución de probabilidad tiene un valor esperado igual al verdadero parámetro que se está estimando.

En adición de saber que en promedio se tiene una estimación correcta, también se desea saber qué tan probable o improbable es estar lejos del verdadero parámetro en una muestra dada. Esto significa que se desea conocer la distribución de b (alrededor de su media β). Primero que todo, la varianza de b condicional en X esta dada por:

$$V\{b|X\} = \sigma^2(X'X)^{-1} = \sigma^2\Big(\sum_{i-1}^{N} x_i x_i'\Big)^{-1} \tag{3.45}$$

que, por simplicidad, se denota como $V\{b\}$. La matriz $K \times K$ $V\{b\}$ es una matriz de varianza convarianzas, que contiene las varianzas de $b_1, b_2, ..., b_K$ en la diagonal, y sus covarianzas como elementos fuera de la diagonal.

Demostración

$$V\{b\} = E\{(b-\beta)(b-\beta)'\} = E\{(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}\} \tag{3.46}$$

$$= (X'X)^{-1}X'(\sigma^2 I_N)X(X'X)^{-1} = \sigma^2(X'X)^{-1} \tag{3.47}$$

Sin usar notacion matricial la demostración es como sigue:

$$V\{b\} = V\Big\{\Big(\sum_i x_i x_i'\Big)^{-1}\sum_i x_i\epsilon_i\Big\} = \Big(\sum_i x_i x_i'\Big)^{-1} V\Big\{\sum_i x_i\epsilon_i\Big\}\Big(\sum_i x_i x_i'\Big)^{-1} \tag{3.48}$$

$$= \Big(\sum_i x_i x_i'\Big)^{-1} \sigma^2 \Big(\sum_i x_i x_i'\Big)\Big(\sum_i x_i x_i'\Big)^{-1} = \sigma^2 \Big(\sum_i x_i x_i'\Big)^{-1} \tag{3.49}$$

Esto requiere los supuestos $S1$-$S4$.

Un último resultado se obtiene del teorema de Gauss Markov, que dice que bajo los supuestos $S1$-$S4$ el estimador de MCO b es el **mejor estimador lineal e insesgado** para β. Se suele decir que b es **MELI** para β. Para apreciar este resultado, considere la clases de estimadores lineales e insesgados. Un estimador lineal es una función lineal de los elementos de y y puede ser reescrito como $b = Ay$, donde A es una matriz $K \times K$. El estimador es insesgado si $E\{Ay\} = \beta$. (Note que el estimador de MCO es obtenido para $A = (X'X)^{-1}X'$). Entonces el teorema establece que la diferencias de matrices de covarianzas de $\tilde{b} = Ay$ y el de MCO es siempre positiva semi definida. ¿Qué esto significa? Suponga que estamos interesados en una combinación lineal de los coeficientes de β, dado por $d'\beta$, donde β es un vector K-dimensional. Entonces, el resultado de Gauss Markov implica que la varianza del estimador de MCO $d'b$ para $d'\beta$ no es mayor que la varianza de otro estimador lineal e insesgado $d'\tilde{b}$, esto es:

$$V\{d'\tilde{b}\} \geq V\{d'b\} \tag{3.50}$$

para cualquier vector d.

Como un caso especial esto se cumple para el elemento k y se tiene

$$V\{\tilde{b}_k\} \geq V\{b_k\} \tag{3.51}$$

Entonces, bajo los supuestos de Gauss Markov, el estimador de MCO es el estimador más preciso, lineal e insesgado para β.

Para estimar la varianza de b se necesita reemplazar la varianza desconocida del error σ^2 con un estimado. Un candidato obvio es la varianza muestral de los residuos $e_i = y0 - x_i'b$, esto es:

$$\tilde{S}^2 = \frac{1}{N-1}\sum_{i=1}^{N} e_i^2 \tag{3.52}$$

Sin embargo, dado que e_i es diferente para ϵ_i, se puede mostrar que es un estimador sesgado para σ^2. Un estimador insesgado esta dado por:

$$s^2 = \frac{1}{N-K}\sum_{i=1}^{N} e_i^2 \tag{3.53}$$

Este estimador tiene una corrección de grados de libertad al dividir por el número de observaciones menos el número de regresores (incluyendo el intercepto). Un argumento intuitivo para esto es que los K parámetros fueron elegidos para minimizar la suma de residuos al cuadrado y por tanto minimizar la varianza muestral de los residuos. Consecuentemente, se espera que $\tilde{s}$ subestime la varianza del término de error σ^2. El estimador s^2, con corrección por grados de libertad, es insesgado bajo los supuestos $S1$-$S4$. La varianza de b puede ser estimada por:

$$V\{b\} = s^2(X'X)^{-1} = s^2\Big(\sum_{i=1}^{N} x_i x_i'\Big)^{-1} \tag{3.54}$$

La varianza estimada de un elemento b_k está dada por $s^2 c_{kk}$, donde c_{kk} es el elemento (k,k) en $(\Sigma_i x_i x_i')^{-1}$. La raíz cuadrada de esta varianza estimada se conoce usualmente como el **error estándar** de b_k. Se puede

denotar como $se(b_k)$. Es la desviación estándar estimada de b_k y es una medida de la precisión del estimador. Bajo los supuestos $S1$-$S4$, se tiene que $se(b_k) = s\sqrt{c_{kk}}$. Cuando los términos de error no son homocedásticos o exhiben autocorrelación, el error estándar del estimador de MCO b_k puede ser computado en una forma diferente.

En general, la expresión para estimar la matriz de covarianza en (3.54) no admite derivación de expresiones analíticas para el error estándar de un solo elemento b_k. Como una ilustración, sin embargo, considerese el modelo de regresión con dos variables explicativas y una constante:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

En este caso es posible derivar la varianza del estimador b_2 para β_2 esta dada por:

$$V\{b_2\} = \frac{\sigma^2}{1 - r_{23}^2} \Big[\sum_{i=1}^{N} (x_{i2} - \bar{x}_2)^2 \Big]^{-1}$$

donde r_{23} es el coeficiente de correlacion muestral entre x_{i2} y x_{i3}, y $\bar{x}_2$ es el promedio muestral de x_{i2}. Se puede reeescribir como:

$$V\{b_2\} = \frac{\sigma^2}{1 - r_{23}^2} \frac{1}{N} \Big[\frac{1}{N} \sum_{i=1}^{N} (x_{i2} - \bar{x}_2)^2 \Big]^{-1}$$

Esto muestra que la varianza de b_2 se determina por 4 elementos. Primero, el término en paréntesis al cuadrado denota la varianza muestral de x_2: Mayor variación en los valores de los regresores llevan a un estimador más preciso. Segundo, el término $\frac{1}{N}$ esta relacionado de manera inversa al tamaño de muestra: más observaciones incrementan la precisión. Tercero, mientras mayor es la varianza del error σ^2, mayor la varianza del estimador. Un valor bajo de σ^2 implica que observaciones cercanas a la línea de regresión, son más faciles de estimar. Finalmente, la varianza es determinada por la correlación entre los regresores. La varianza de b_2 está inflada por la correlación entre x_{i2} y x_{i3} es alta (positiva o negativa). En el extremo cuando $r_{23} = 1$ o -1, x_{i2} y x_{i3} están perfectamente correlacionadas y la varianza se vuelve infinitamente grande. Este es el caso de perfecta colinealidad, y el estimador de MCO no puede ser computado.

Los supuestos $S1$-$S4$ establecen que los términos de error ϵ están mutuamente no correlacinados, son independientes de X, tienen media cero y tienen varianza constante, pero no especifican la forma de la distribución. Para la inferencia estadística exacta de una muestra dada de N observaciones, se tienen que hacer supuestos explícitos sobre la distribución. El supuesto más común es que los errores se distribuyen como una normal conjunta. En este caso la no correlación es equivalente a la independencia de todos los términos de error. El supuesto preciso es:

$$\epsilon \sim N(0, \sigma^2 I_N) \tag{S5}$$

que dice que el vector de términos de error ϵ sigue una distribución normal N-variada con vector de medias 0 y matriz de varianza covarianza $\sigma^2 I_N$. El supuesto $S5$ reemplaza $S1$, $S3$ y $S4$. Una forma alternativa de formamular a $S5$ es:

$$\epsilon_i \sim NID(0, \sigma^2)$$

que es una abreviación de decir que los términos de error son extracciones independientes de una distribucióon normal (normal e independientemente distribuidos, o NID) con media cero y varianza σ^2. A pesar que los términos error son no observados, esto no significa que se esta libre de hacer el supuesto que se quiera. Por ejemplo, si los términos de error se asumen que siguen una distribución normal, esto significa que y_i (para valores dados de x_i) también sigue una distibrución normal. Claramente, se puede pensar en muchas

variables cuya distribución no es normal, en cuyo caso el supuesto de términos de error normal es inapropiado. Afortunadamente, no todos los supuestos no son igualmente cruciales para la validez de los resultados que siguen y, más a un, la mayoría de los supuestos pueden ser contrastados empiíicamente.

Para hacer las cosas mas simple, considerese la matriz X como fija y determinística o, alternativamente, condicional en X. Entonces el siguiente resultado se cumple. Bajo los supuestos $S2$ y $S5$ el estimador de MCO b se distribuye normalmente con media β y matriz de covarianzas $\sigma^2(X'X)^{-1}$, esto es,

$$b \sim N(\beta, \sigma^2(X'X)^{-1})$$

Demostración:

Este resultado implica que cada elemento de b esta normalmente distirbuido, por ejemplo:

$$b_k \sim N(\beta_k, \sigma^2 c_{kk})$$

Estos resultados proveen la base para los contrastes estadisticos basados en el estimador de MCO b.

Ejemplo

El problema de Inferencia

Introducción

En este capítulo abordamos el denominado **problema de inferencia**. En el campo de la estadística este abarca tres dimensiones:

1. Estimación de los parámetros desconocidos.
2. Contraste o (testing) de hipótesis económicas
3. Predicción de los resultados del modelo.

En clases anteriores se trató la primera de estas dimensiones, en lo que denominamos el problema de estimación, utilizando los estimadores de mínimos cuadrados para generar estimaciones puntuales de los parámetros del MRL.

Estas estimaciones representan una **inferencia** acerca de la función de regresión:

$$E(y) = X'\beta$$

Comprender el concepto de inferencia es clave para entender del objetivo y el alcance de la econometría. En particular, al *inferir* se están extrayendo por razonamiento conclusiones de algo en base a supuestos y conocimiento.

En nuestro caso, se establecen supuestos acerca del modelo de regresión. Basados en estos supuestos, y dadas las estimaciones de los parámetros, se quieren hacer inferencias acerca de la población de donde provinieron los datos.

En lo que sigue, se evaluarán dos herramientas de inferencia estadísticas adicionales:

1. Estimación de intervalos
2. Contrastes de hipótesis

En el primer caso, se trata de una herramienta para crear rangos de valores, en los que el parámetro desconocido es probable que se encuentre. En el segundo, consiste en procedimientos para comparar conjeturas acerca de los parámetros de regresión, utilizando datos. Ambos procedimientos dependen del supuesto de normalidad.

Estimación de intervalos.

En nuestro ejemplo sobre los determinantes del salario, la estimación de la función de regresión fue:

$$salario = -2.2 + 1.8\ educ$$

Donde $b_1 = -2.2$ y $b_2 = 1.8$ son **estimaciones puntuales** de los parámetros poblacionales desconocidos β_1 y β_2.

Recordemos que si se cumplen los supuestos para el MRL, los estimadores tenían una distribución normal. Por ejemplo, para el caso de b_2,

$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

Sustrayendo la media y dividiendo por la desviación estándar, se puede transformar en una variable normal estándar:

$$Z = \frac{b_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}} \sim N(0,1)$$

Utilizando la tabla de probabilidades normales, sabemos que

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Es decir, la probabilidad de que Z se encuentre en dicho rango es 95%. Susituyendo a Z:

$$P\left(-1.96 \leq \frac{b_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}} \leq 1.96\right) = 0.95$$

Reorganizando,

$$P\left(b_2 - \frac{1.96\sqrt{\sigma^2}}{\sum(x_i - \bar{x})^2} \leq \beta_2 \leq b_2 + \frac{1.96\sqrt{\sigma^2}}{\sum(x_i - \bar{x})^2}\right)$$

Esta expresión define un intervalo que tiene, probabilidad de 0.95 de contener el parámetro β_2. Los dos puntos definen un estimador del intervalo. Dados los supuestos, si se toman múltiples muestras y se estiman los intervalos para cada una de estas muestras, 95% de estos intervalos contendrían el valor del verdadero parámetro.

Note que la derivación depende la varianza del error. En la práctica se trabaja con el estimador de esta que derivamos en el capítulo anterior.

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i}{N-2}$$

donde: $\hat{e}_i = y_i - b_1 - b_2 x_i$

Reemplazado σ^2 por $\hat{\sigma}^2$ en la definición de la variable Z, genera una variable aleatoria con distribución t con $N-2$ grados de libertad,

$$t = \frac{b_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}} = \frac{b_2 - \beta_2}{\sqrt{\hat{var}(b_2)}} = \frac{b_2 - \beta_2}{se(b_2)} \sim t_{N-2}$$

Este resultado también aplica para b_1, por lo que en general,

$$t = \frac{b_k - \beta_k}{se(b_k)} \sim t_{N-K}$$

Donde K es el número de parámetros estimados. Para el caso de m-grados de libertad, el percentil 95 de la distribución t se denota

$$t_{0.95,m}$$

A los valores de "tabla" se les conoce como "valores críticos" o t_c. Provienen de una distribución t tal que sirven para evaluar:

$$P(t \geq t_c)$$

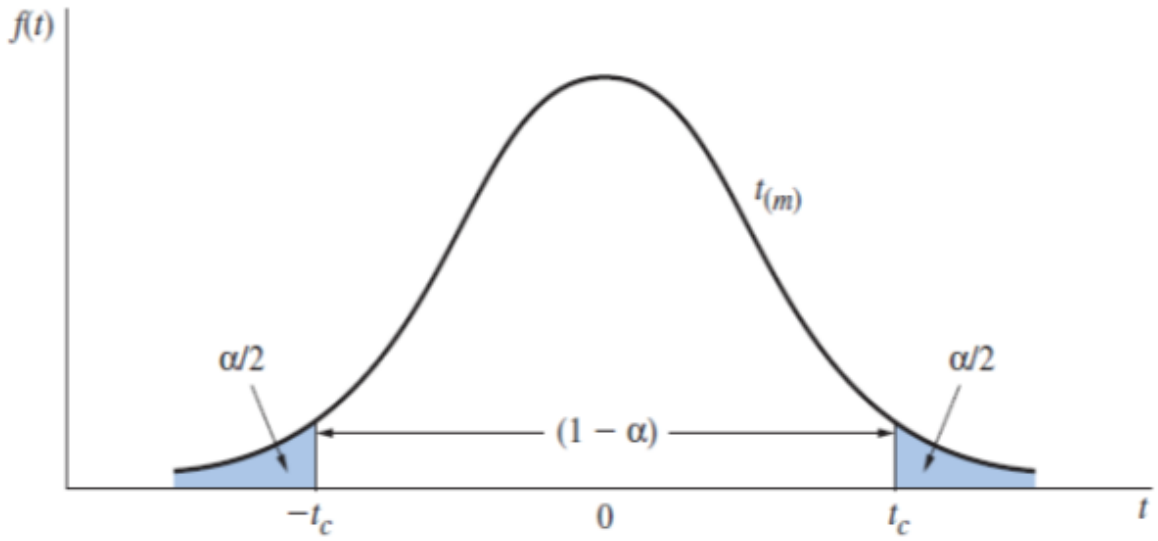
Como la distribucion es simétrica

$$P(t \leq -t_c)$$

Lo que se trata de buscar es el valor crítico tal que:

$$P(t \geq t_c) = P(t \leq t_c) = \frac{\alpha}{2}$$

Donde α es la probabilidad, por lo general es 0.01 ó 0.05. La figura 1 muestra los valores críticos.



El intervalo de confianza del modelo de regresión lineal con $m = N - 2$ grados de libertad es,

$$P(-t_{(0.975,N-2)} \leq t \leq t_{(0.975,N-2)})$$

Sustituyendo la definición de t en el intervalo de confianza:

$$P\left(-t_c \leq \frac{b_k - \beta_k}{se(b_k)} \leq t_c \right)$$

Reorganizando, se obtiene una expresión para estimar el intervalo:

$$P\left(b_k - t_c se(b_k) \leq \beta_k \leq b_k + t_c se(b_k) \right) = 1 - \alpha$$

Los extremos del intervalo, $b_k - t_c se(b_k)$ y $b_k + t_c se(b_k)$ son aleatorios, pues cambian de muestra en muestra. Define un **estimador de intervalo** para β_k. Cuando se calcula con una muestra en particular se denomina **estimado del intervalo** de β_k. Otra manera de llamarlo es **intervalo de confianza**.

Interpretación:

Si seleccionan muchas muestras de tamaño N, donde para cada una de estas se computa b_k y $se(b_k)$, y luego la estimación del intervalo $b_k \pm t_c se(b_k)$, entonces $100(1-\alpha)\%$ de todos los intervalos construidos pueden contener el verdadero β_k. Un intervalo en particular puede contener o no a β_k siendo esto desconocido por la naturaleza de β_k. Luego, el término **intervalo de confianza** se refiere al procedimiento parsa construir el intervalo y no a una estimación particular del intervalo.

Retomando el ejemplo anterior, supongamos que disponemos de los siguientes datos para estimar la relación entre el salario y la educación. Nos piden estimar el siguiente modelo de regresión:

$$salario_i = \beta_1 + \beta_2 educ_i + \epsilon_i$$

Obs	Salario (miles de RD$)	Educ (Años)
1	10	5
2	15	12
3	10	8
4	30	16
5	18	12

Aplicando las fórmulas, los resultados son: $b_1 = -2.2$ y $b_2 = 1.8$. Una forma conveniente de presentar los resultados es en términos de la función de regresión muestral:

$$\hat{y}_i = -2.2 + 1.8 x_i$$

En este caso $N = 5$, por lo que los grados de libertad (gl) son:

$$gl = N - 2 = 5 - 2 = 3$$

Para un intervalo de confianza de 95%, $\alpha = 0.05$. El valor crítico $t_c = t_{(1-\frac{\alpha}{2},N-2)} = t_{0.975,3} = 3.182$

AQUI TABLA VALORES CRITICOS

Por lo que para el caso de la pendiente β_2,

$$P\Big(b_2 - 3.182se(b_2) \leq \beta \leq b_2 + 3.182se(b_2)\Big)$$

Para computar $se(b_2)$:

$$se(b_2) = \sqrt{var(b_2)} = \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{15.8}{71.2}} = 0.4711$$

Donde:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}^2}{N-2} = 15.8$$

$$\sum(x_i - \bar{x})^2 = 71.2$$

Sustituyendo se obtiene el estimado del intervalo de confianza al 95% para β_2

$$b_2 \pm t_c se(b_2) = 1.8 \pm 3.182(0.4711) = [0.30, 3.30]$$

Es decir, con un "95% de confianza" por cada año de educación adicional, el salario promedio incrementa entre 300 y 3300 pesos.

Una nota final. No sabemos si β_2 está en este intervalo, y nunca lo sabremos. Lo que si sabemos es que cuando el procedimento sea utilizado en muchas muestras aleatorias de la misma población, en 95% de todos los intervalos construidos contendrán el verdadero parámetro. Lo que se puede decir de este intervalo, es que estaríamos sorprendidos si el parámetro no estuviera en él.

Cuál es la ventaja de este procedimiento en comparación a la estimación puntual?

1. La estimación puntual no provee información sobre su viabilidad. Por lo que hay que reportar su intervalo.
2. Incorpora tanto la estimación puntual como la desviación estándar.
3. Incorpora el impacto del tamaño de muestra. Mientras menor la muestra, más ancho es el intervalo y más impreciso la inferencia acerca de β_2.
4. Si el intervalo es muy ancho, sirve para tomar la decisión de si es necesaria una nueva muestra o más datos para ser concluyente acerca de los resultados.

Contraste de hipótesis

El propósito del contraste de hipótesis consiste en verificar:

1. Si las proposiciones de la teoría son avaladas por un conjunto de datos específicos.
2. Si un determinado parámetro es un valor específico.

Formalmente, los procedimientos de contraste de hipótesis comparan una conjetura acerca de una población con la información contenida en una muestra.

Los componentes de los contrastes (test) de hipótesis son:

1. Una hipótesis nula H_0
2. Una hipótesis alternativa H_1
3. Un contraste estadístico
4. Una región de rechazo
5. Una conclusión

Hipótesis Nula (H_0)

Es la creencia que se mantiene hasta que la evidencia muestral arroja que es falsa. Cuando eso sucede, se dice que se rechaza. Es especifica el valor para un parámetro de la regresión. Se formula como:

$$H_0 : \beta_k = c$$

donde c es una constante.

Hipótesis alternativa (H_1)

Es la hipótesis a aceptar cuando la nula es rechazada y depende de la teoría económica. Para $H_0 : \beta_k = c$, hay tres posibles hipótesis alternativas:

1. $H_1 : \beta_k > c$: rechazar la H_0 resulta en aceptar la alternativa que $\beta_k > c$. En economía se utiliza mucho la desigualdad en la hipótesis alternativa, pues la teoría suele solo sugerir el signo.
2. $H_1 : \beta_k < c$: rechazar la H_0 resulta en aceptar la alternativa que $\beta_k < c$.
3. $H_1 : \beta_k \neq c$: rechazar la H_0 resulta en aceptar la alternativa que β_k toma un valor mayor o menor que c.

Estadístico de contraste

La información muestral acerca de la H_0 est| incluida en el valor muestral de un contraste estadístico. Basado en este estadístico, se decide si rechazar o no la hipótesis nula.

Una característica del contraste estadístico es que la distribución de probabilidad es conocida si la H_0 es verdadera, mientras que si no es verdadera no se conoce.

Si la $H_0 : \beta_k = c$ es cierta, entonces se utiliza el estadístico t que vimos:

$$t = \frac{b_k - c}{se(b_k)} \sim t_{N-2}$$

Si la nula no es cierta, entonces t no tiene una distribución t con $N - 2$ grados de libertad.

La zona o región de rechazo

Es el rango de valores del contraste estadístico que lleva al rechazo de la hipótesis nula. Se construye a partir de la siguiente información:

1. Un contraste estadístico cuya distribución es conocida cuando la nula es cierta.
2. Una hipótesis alternativa
3. Un nivel de significancia o probabilidad de rechazar la nula cuando es cierta.

Esta zona consiste en valores que son poco probables o tienen poca probabilidad de ocurrir cuando la nula es cierta. Es decir, si el valor del estadístico de contraste se ubica en dicha zona, es poco probable que esa sea la distribución de este estadístico, y en consecuencia que la H_0 sea cierta.

Otra forma de verlo, es que si la H_1 es cierta entonces el valor del test estadístico va a ser "muy grande" o "muy pequeño". Donde el nivel de significancia es que determina cuán grande o pequeño sea un valor.

Obviamente, cuando rechazamos una H_0 existe una probabilidad mayor a cero que sea cierta. En ese caso estamos cometiendo Error Tipo I. Este error es medible y está representado por α. Es decir, se puede escoger el grado de Error Tipo I.

El otro tipo de error es el Error Tipo II. Este surge cuando no rechazamos una H_0 que es falsa. No es medible, pues depende del parámetro desconocido β_k.

La zona o región de rechazo para distintas alternativas:

1. Contraste de una cola cuando la alternativa es "mayor a ($>$)":

En este caso el t calculado tiene que ser mayor que el t crítico (de tabla). Por lo que se rechaza la nula dado un nivel de significancia.

AQUI FIG

2. Contraste de una cola cuando la alternativa es "menor a ($<$)":

En este caso el t calculado tiene que ser menor que el t crítico (de tabla). Por lo que se rechaza la nula dado un nivel de significancia.

AQUI FIG

3. Contraste de una cola cuando la alternativa es "no igual a ($\neq$)"

En este caso el t calculado tiene que ser menor o mayor que el t crítico (de tabla). Por lo que se rechaza la nula dado un nivel de significancia $\alpha/2$.

AQUI FIG

Procedimiento para contraste de hipótesis

1. Determine la H_0 y las hipótesis alternativas.
2. Especifique el contraste estadístico y su distribución si la nula es verdadera.
3. Seleccione α y determine la zona de rechazo.
4. Calcule el valor muestral del estadístico.
5. Establezca la conclusión.

Valor p o p-value

Se refiere al **valor de la probabilidad** de obtener un resultado al menos tan extremo como el que realmente se ha obtenido, suponiendo que la H_0 es cierta. Por ejemplo, en el caso de un contraste de cola derecha:

$$p = P(t_m \geq t)$$

Para fines de determinar el resultado de un contraste, el valor - p se compara con el nivel de significancia escogido. Si $p < \alpha$ se rechaza la H_0. Si $p > \alpha$ no se rechaza la H_0.

Ejemplos

Contraste de hipótesis en el MRL múltiple

En adición al contraste de hipótesis para un solo parámetro, otros contrastes son frecuentes en las aplicaciones econométricas. Estas son las llamdas pruebas de restricciones lineales múltiples, las cuales se basan en la prueba F.

Estas pruebas consisten en evaluar la hipótesis de si un conjunto de variables independientes no tiene efecto parcial sobre la variable dependiente. Se pueden clasificar en:

1. Pruebas de restricciones de exclusión
2. Prueba de significancia general de una regresión
3. Prueba de restricciones generales lineales

Pruebas de restricciones de exclusión.

Se desea probar si un grupo de variables no tiene efecto sobre la variable dependiente. Considere el siguiente modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Suponga que se tiene interés en evaluar si x_3, x_4 y x_5 no tienen efectos sobre y. La hipótesis nula es:

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$

Las cuales se denominan **restricciones de exclusión**. En este ejemplo se dice que se están evaluando 3 restricciones de exclusión. También se conocen como **restricciones múltiples**.

A las pruebas de hipótesis de este tipo se les denomina **prueba de hipótesis múltiple o conjunta**.

La hipótesis alternativa es:

$H_1 : H_0$ no es verdadera.

Es decir, se satisface para algún β_j en la H_0 que sea diferente de cero.

Cómo probar estas hipótesis?

Pasos:

1. Estime el modelo sin las restricciones. Este modelo le denominamos **modelo no restringido (MNR)**. Guarde la suma de residuos al cuadrado de este modelo como SRC_{MNR}.
2. Estime el modelo con las restricciones: **modelo restringido (MR)**. Guarde la suma de residuos al cuadrado de este modelo como SRC_{MR}.
3. Compute el siguiente estadístico de prueba:

$$F = \frac{(SRC_{MR} - SRC_{MNR})/q}{SRC_{MNR}/(N-K)}$$

Donde $q = gl_{MR} - gl_{MNR}$

La distribución de este estadístico es:

$$F \sim F_{q,N-K}$$

También se puede computar usando las R^2 de ambos modelos:

$$F = \frac{(R^2_{MNR} - R^2_{MR})/q}{(1 - R^2_{MNR})/(N - K)}$$

Prueba de significancia general de una regresión

En este caso la hipótesis es que las variables del modelo no explican a la variable dependiente:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_K$$

Note que sería como imponer las K restricciones lineals. Por lo que el modelo restringido es:

$$y = \beta_0 + \epsilon$$

El estadístico F en este caso es:

$$F = \frac{R^2/K}{(1 - R^2)/(N - K)}$$

Donde R^2 es la relativa al modelo no restringido, es decir, al que tiene todos los regresores. A este estadístico también se le conoce como **significancia conjunta de la regresión**.

Prueba de restricciones generales lineales

El estadístico F discutido, sirve también para probar hipótesis relativas a una teoría. Por ejemplo, la teoría de la producción sugiere que el nivel de producto en una industria depende del trabajo y el capital físico:

$$y = \beta_0 + \beta_1 l + \beta_2 k + \epsilon$$

Donde y es el nivel de producción, l es el nivel de trabajo contratado y k es el nivel de capital físico contratado.

Una hipótesis usual es la de rendimientos constantes a escala. Es decir,

$$H_0 : \beta_1 + \beta_2 = 1$$

En este caso el modelo restringido es:

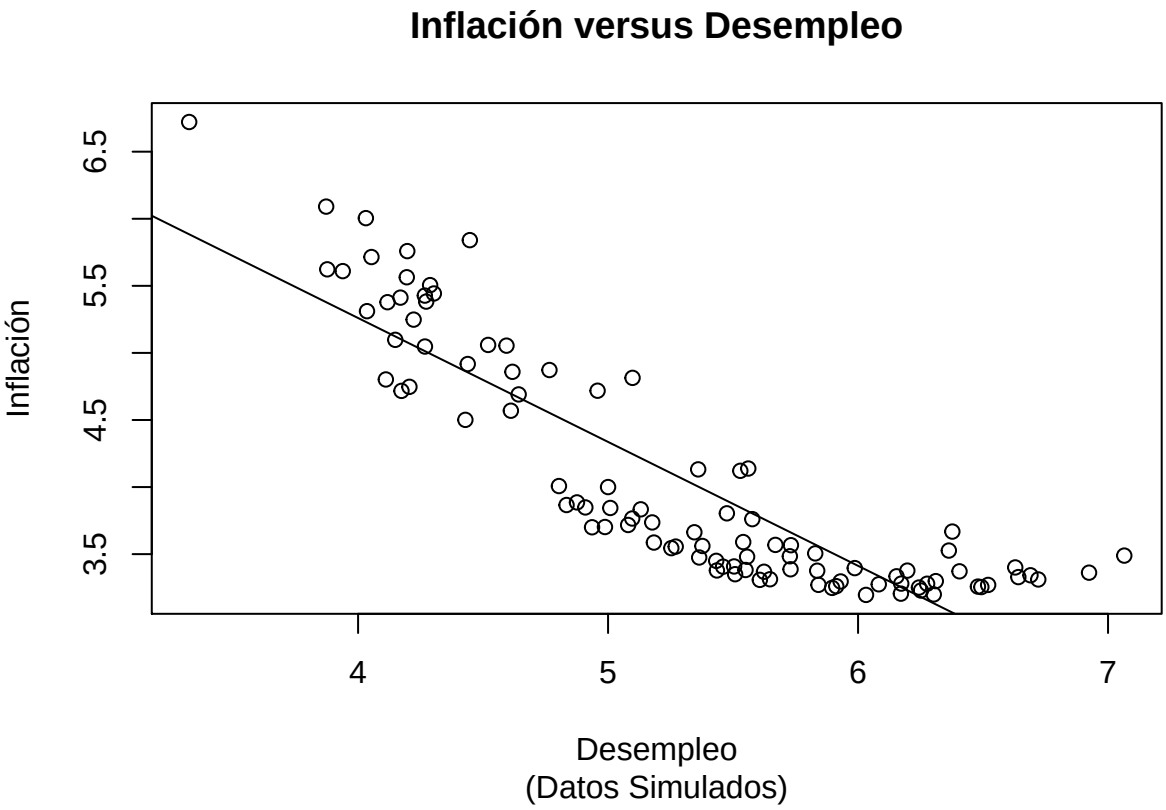
$$y = \beta_0 + \beta_1 l + (1 - \beta_1)k + v$$

Y la prueba es una F que compara las SRC entre el modelo restringido y no restringido.

Forma funcional

Introducción

Hasta ahora hemos asumido una forma funcional lineal para la función de regresión poblacional. Sin embargo, en la práctica las relaciones entre las variables de interés pueden ser o no lineales. Incluso, hay situaciones donde la teoría puede sugerir relaciones no lineales. Por ejemplo, la existencia de rendimientos crecientes (o decrecientes).



La figura anterior muestra la relación entre la inflación y desempleo. De acuerdo a la teoría macroeconómica se espera que mayor desempleo menores niveles de inflación, es decir, una relación negativa. La linea con pendiente negativa corresponde a la especificación de dicha relación a través del modelo de regresión lineal:

$$inflacion_i = \beta_0 + \beta_1 desempleo_i + \epsilon_i$$

Se puede apreciar que el ajuste que hace dicho modelo de los datos es mediocre, en el sentido que para niveles bajos de desempleo, el modelo genera sistamáticamente predicciones por debajo de los datos observados y para niveles altos lo contrario. Es decir, la función de regresión muestral no representa de manera adecuada la relación entre estas variables. Para subsanar este problema, se requiere una especificación que permita que el efecto de la tasa de desempleo sobre la inflación cambie con el nivel de desempleo. Es decir, un modelo que capture ese comportamiento no lineal destacado. Una especificación que cumplen dicho propósito son:

$$inflacion_i = \beta_0 + \beta_1 desempleo_i + \beta_2 desempleo_i^2 + \epsilon_i$$

con $\beta_1 > 0$ y $\beta_2 < 0$.

El efecto parcial es:

$$\frac{\partial inflacion_i}{\partial desempleo} = \beta_1 + 2\beta_2 desempleo$$

Note que, dado los signos mencionados, el efecto marginal del desempleo depende (negativamente) del nivel de desempleo. Más aún, a medida que el desempleo aumenta, menor es la respuesta de la inflación.

Este ejemplo ilustra que el modelo de regresión lineal puede acomodar relaciones no lineales para capturar mejor el comportamiento de las variables bajo análisis. Esto es así, porque el supuesto de linealidad se refiere a que el modelo sea lineal en los parámetros, de tal manera que la función de regresión poblacional pueda ser estimada por el método de mínimos cuadrados ordinarios.

Asimismo, existen situaciones donde la relación sugerida por la teoría es no lineal (en los parámetros), pero bajo una transformación adecuada, el modelo pueda ser reescrito como un modelo lineal en los parámetros. Por ejemplo, considere la siguiente función de producción Cobb-Douglas:

$$Y_i = A_i K_i^{\beta_1} L_i^{\beta_2}$$

Donde $A_i = \bar{A} exp(\epsilon_i)$ es la denominada productividad total de factores.

Este modelo es no lineal ni en los parámetros, por lo que no puede estimarse por MCO. Sin embargo, aplicando logarimo (natural):

$$\ln Y_i = \ln A_i + \beta_1 \ln K_i + \beta_2 \ln L_i$$

$$\ln Y_i = \ln \bar{A} + \beta_1 \ln K_i + \beta_2 \ln L_i + \epsilon_i$$

$$\ln Y_i = \beta_0 + \beta_1 \ln K_i + \beta_2 \ln L_i + \epsilon_i$$

Se obtiene una especificación cuyos parámetros pueden ser estimados por MCO.

Por otro lado, algunos modelos no lineales no pueden ser transformados en lineales. Por ejemplo, el modelo:

$$y_i = \beta_0 + x_i^{\beta_1} + \epsilon_i$$

es obvio que no puede ser transformado y no estimable por MCO pues el parámetro β_1 entra de manera no lineal en la especificación.

Otros ejemplos serían:

a) $y_i = \beta_0 k_i^{\beta_1} l_i^{\beta_2} + \epsilon_i$

b) $y_i = (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})\epsilon_i$

Formas funcionales útiles

Algunas forms funcionales no lineales son utilizadas de manera recuerrente en las aplicaciones econométricas (las cuales se ilustran con el modelo de regresión simple por fines de conveniencia expositiva) son las siguientes:

1. Modelo log-log

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + \epsilon_i$$

2. Modelo log-lin

$$\ln y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

3. Modelo lin-log

$$y_i = \beta_0 + \beta_1 \ln x_i + \epsilon_i$$

4. Cuadrático

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

5. Cúbica

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$

6. Recíproca

$$y_i = \beta_0 + \beta_1 \frac{1}{x_i} + \epsilon_i$$

Note que todos son lineales en los parámetros y por tanto estimables por MCO.

Interpretación de los efectos parciales como elasticidades

Una forma conveniente de interpretar los efectos parciales es como elasticidades y semi-elasticidades, puesto que su comunicación es independiente de las unidades de las variables del modelo y además se pueden estimar directamente de algunas de las especificaciones arriba mencionadas.

El concepto de elasticidad se refiere al cambio porcentual de una variable Y respecto como resultado del cambio porcentual de una variable X. Por ejemplo, en los cursos de introducción a la economía suele estudiarse la denominada elasticidad precio de la demanda, definida como:

$$\eta_{QP} = \frac{\Delta Q}{\Delta P}\frac{P}{Q} \approx \frac{\Delta \ln Q}{\Delta \ln P}$$

usualmente multiplicada por 100, para su interpretación en términos porcentuales.

Partiendo de esa definición podemos obtener en términos de elasticidades los efectos parciales de algunas de las especificaciones para el MRL:

a) Modelo lineal: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

- Efecto parcial:

$$\frac{\Delta y_i}{\Delta x_i} = \beta_1$$

- Elasticidad: multiplicar el efecto parcial por $\frac{x_i}{y_i}$

$$\frac{\Delta y_i}{\Delta x_i}\frac{x_i}{y_i} = \beta_1 \frac{x_i}{y_i}$$

b) Modelo log-log: $\ln y_i = \beta_0 + \beta_1 \ln x_i + \epsilon_i$

- Efecto parcial:

$$\frac{\Delta \ln y_i}{\Delta \ln x_i} = \beta_1$$

- Elasticidad:

$$\frac{\Delta \ln y_i}{\Delta \ln x_i} = \beta_1$$

c. Modelo log-lin: $\ln y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

- Efecto parcial:

$$\frac{\Delta \ln y_i}{\Delta x_i} = \beta_1$$

- Elasticidad:

$$\frac{\Delta \ln y_i}{\Delta x_i} x_i = \beta_1 x_i$$

c. Modelo lin-log: $y_i = \beta_0 + \beta_1 \ln x_i + \epsilon_i$

- Efecto parcial:

$$\frac{\Delta y_i}{\Delta \ln x_i} = \beta_1$$

- Elasticidad:

$$\frac{\Delta y_i}{\Delta \ln x_i} \frac{1}{y_i} = \beta_1 \frac{1}{y_i}$$

Note que la elasticidad en estas especificaciones (excepto el modelo log-log) no son constantes, sino que varía para cada individuo. Para obtener una sola estimación de la elasticidad de interés se pueden calcular el efecto parcial promedio (EPP) como el valor promedio de la variable y sustituirlo en el efecto de interés. Por ejemplo, en el modelo log-lin es: $EPP = \beta_1 \bar{x}$, donde $\bar{x}$ es la media muestral de x_i.

Violación de Supuestos del MRL

Introducción

La caracterización empírica de la relación causal sugerida por la teoría económica, fue plasmada en el modelo de regresión lineal. La especificación de este modelo se produjo estableciendo supuestos sobre los distintos elementos que lo componen: las variables observables (dependiente y explicativas), su relación y el componente no observable o término error. Estos supuestos eran:

1. El modelo es lineal en los parámetros y el error entra de forma aditiva.

$$y_i = x_i' \beta + \epsilon_i$$

2. $E(\epsilon_i | x_i) = 0$
3. Homocedasticidad: $var(\epsilon_i | x_i) = E[\epsilon_i^2 | x_i] = \sigma^2$
4. No correlación serial (no autocorrelación): $E[\epsilon_i \epsilon_j | x_i] = 0$
5. No multicolinealidad: x_{ik} no son funciones exactas de otras variables
6. $\epsilon_i \sim N(0, \sigma^2)$

Estos supuestos permitían la especificación de la función de regresión poblacional como una función lineal del conjunto de variables explicativas, la obtención de un estimador de los parámetros desconocidos cuyas propiedades incluyen insesgamiento y varianza mínima (expresadas en el teorema Gauss-Markov) y la posibilidad de hacer inferencia sobre los coeficientes estimados.

En esta sección se analiza el impacto que tiene sobre la estimación e inferencia de los parámetros del MRL, la violación de los supuestos sobre los cuales se formuló el MRL. En particular, el análisis se realiza alrededor de tres preguntas:

1. Cuáles son las consecuencias de la violación de un determinado supuesto sobre las propiedades del estimador de MCO?
2. Cómo se puede detectar la inviabilidad o la violación de un supuesto?
3. Cuáles alternativas existen para subsanar los efectos de la violación de uno de los supuestos del MRL?

Heterocedásticidad

Hasta ahora hemos asumido que el valor promedio de la variable y es explicado por x a través de una función lineal

$$E(y_i|x_i) = x_i'\beta$$

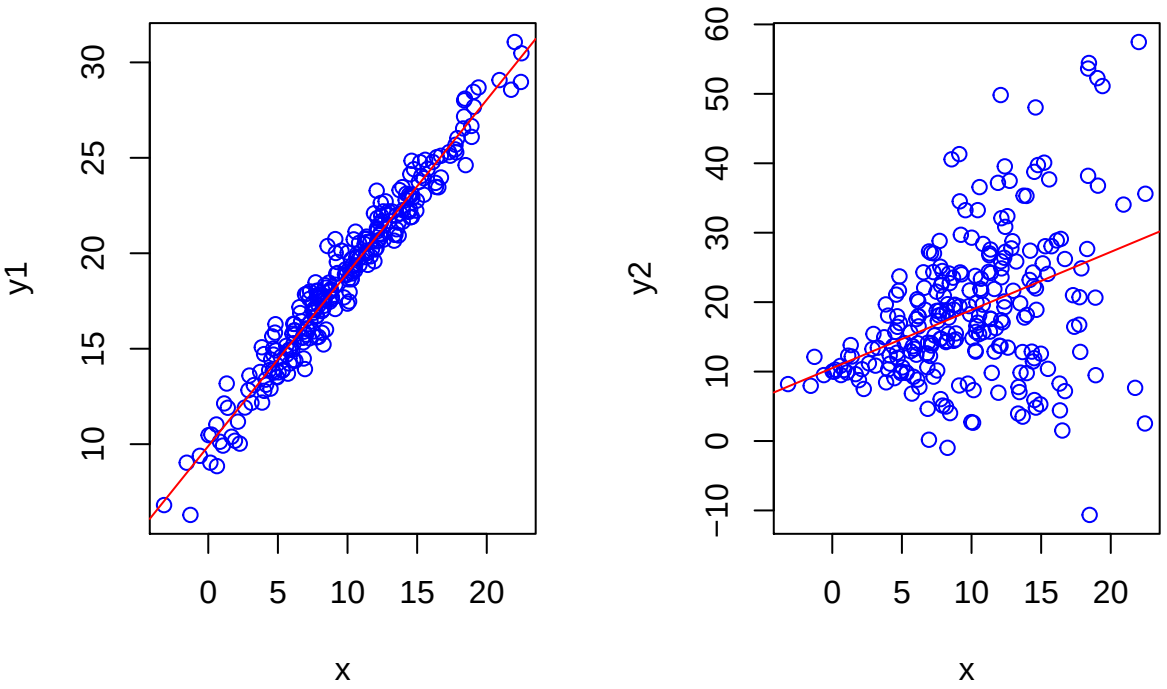
Para reconocer el impacto de otras variables no considerada, se incluye la variable ϵ_i para tomar encuenta esos factores:

$$\epsilon_i = y_i - E(y_i|x_i) = y_i - x_i'\beta$$

Por lo que el modelo para describir el comportamiento de y_i es:

$$y_i = x_i'\beta + \epsilon_i$$

Entre los supuestos en torno al modelo esta el de **homocedasticidad** o igual varianza para todos los ϵ_i. Es decir, que la probabilidad de observar un valor muy grande (o muy pequeño) de ϵ_i es casi la misma entre todos los individuos y submuestras. Un resultado es que se obtiene un estimador de la varianza de los errores, asumiento que es representativo para todos los individuos. Sin embargo, hay situaciones donde este supuesto no es conveniente, pues puede llevar a conclusiones errada de la relación entre las variables. El gráfico siguiente muestra un ejemplo donde este supuesto es válido (izquierda) y uno donde no es válido (derecha).



Note que (en el gráfico de la derecha) se distinguen dos grupos de datos en esta muestra. El primer grupo tiene una dispersion menor que el segundo. En esta situación el supuesto de heterocedasticidad es inadecuado. Los errores del modelo se dice que en este caso son **heterocedásticos**.

Gráficamente, lo que destaca es que la probabilidad de observar errores relativamente extremos es mayor entre distinto grupo de observaciones. En este gráfico ilustra el modelo de regresión lineal en presencia de heterocedasticidad.

Se procede a reemplazar el supuesto:

$$var(\epsilon_i|x_i) = \sigma^2$$

por

$$var(\epsilon_i|x_i) = h(x_i)$$

Donde $h(x_i)$ refleja que la varianza de ϵ_i es una función de x_i.

Es importante destacar que la heterocedásticidad es un problema encontrado comunmente cuando se utilizan datos de corte transversal.

Consecuencias de la heterocedasticidad Los estimadores de MCO siguen siendo lineales e insesgados. Para ver esto note que:

$$b = (X'X)^{-1}Xy = \beta + (X'X)^{-1}X\epsilon$$

Aplicando valor esperado:

$$E(b|X) = \beta + (X'X)^{-1}X'E(\epsilon|X)$$

Dado el supuesto de exogeneidad, se tiene que:

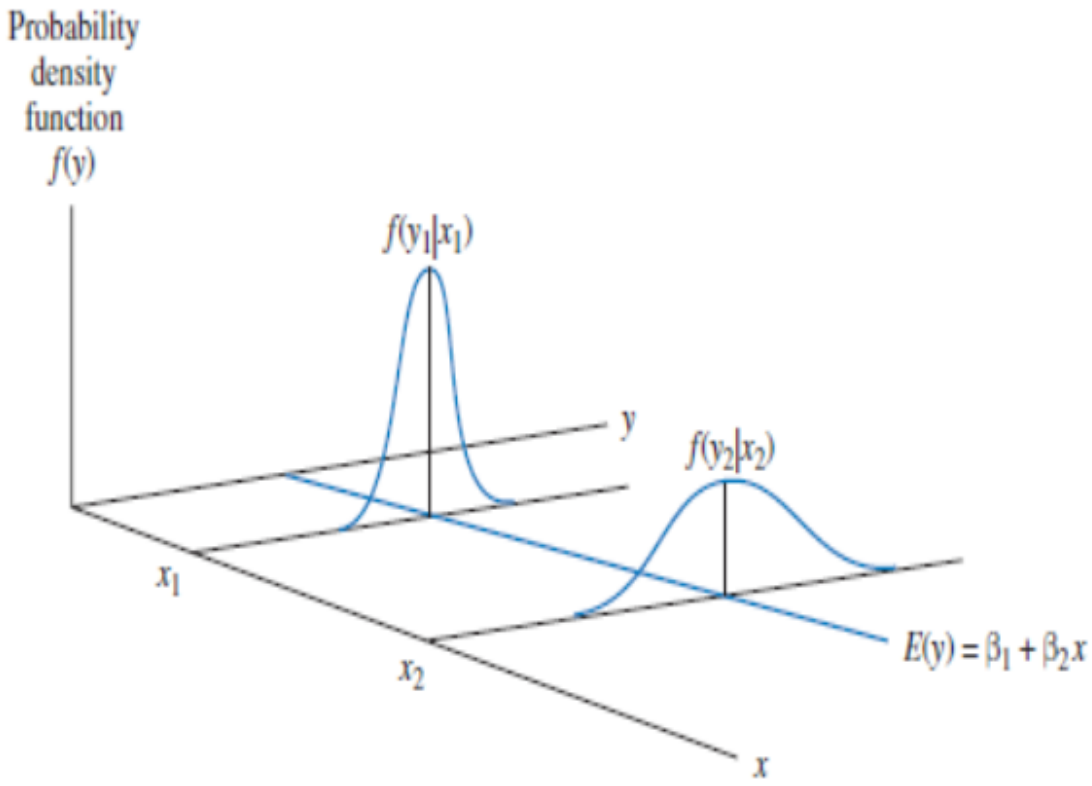


Figure 5: Ejemplo: Distribución Heterocedástica

$$E(b|X) = \beta$$

Es decir, el insesgamiento no se ve afectado por la presencia de heterocedásticidad.

Sin embargo, la matriz de covarianzas (y por tanto los errores estándar), $var[b|X] = \sigma^2(X'X)^{-1}$ es la incorrecta. Para ver esto note que:

$$\begin{aligned} var[b|X] &= E[(b-\beta)(b-\beta)'|X] = E\{[(X'X)^{-1}X'\epsilon][(X'X)^{-1}X'\epsilon]'|X) \\ &= E\{[(X'X)^{-1}X'\epsilon][\epsilon'X(X'X)^{-1}]'|X) \end{aligned}$$

$$= (X'X)^{-1}X'E(\epsilon\epsilon'|X)X(X'X)^{-1}$$

Definiendo como $E(\epsilon\epsilon'|X) = \Omega_\epsilon$ como la matriz de covarianzas, que tienen sobre la diagonal σ_i^2 (heterocedásticidad) para $i = 1, ..., N$ y ceros fuera de la diagonal (no correlación serial), se tiene que la matriz de varianzas de b es distinta al caso homocedástico.

La implicancia es que los errores estándar usandos en la etapa de inferencia, es decir, en las pruebas de hipotesis e intervalos de confianza estan sesgados y toda la inferencia no es válida.

En el caso del modelo de regresión simple, puede apreciarse mejor el impacto de este problema. Como se vio en capítulos pasados la varianza para el caso de la pendiente se obtiene de:

$$b_2 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$var(b_2|x_i) = E[(b_2 - \beta_2)^2|x_i] = E\left[\left(\frac{\sum_i (x_i - \bar{x})\epsilon_i}{\sum_i (x_i - \bar{x})^2}\right)^2 \middle| x_i\right]$$

$$var(b_2|x_i) = E\left[\frac{\left(\sum_i (x_i - \bar{x})\epsilon_i\right)^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \middle| x_i\right]$$

Analizando el numerador, tenemos:

$$E\left[\left(\sum_i (x_i - \bar{x})\epsilon_i\right)^2 \middle| x_i\right] = E\{[(x_1 - \bar{x})\epsilon_1 + ... + (x_N - \bar{x})\epsilon_N]^2|x_i\}$$

$$= E\{[(x_1 - \bar{x})^2\epsilon_1^2 + ... + (x_N - \bar{x})^2\epsilon_N^2 + terminos\ cruzados]|x_i\}$$

Donde no se desarrollan los términos cruzados, que dado el supuesto de no correlación serial, son todos igual a cero cuando se aplica la esperanza matemática. Por tanto,

$$= (x_1 - \bar{x})^2 E(\epsilon_1^2|x_1) + ... + (x_N - \bar{x})^2 E(\epsilon_N^2|x_N) = \sum_i (x_i - \bar{x})^2 E(\epsilon_i^2|x_i)$$

Reemplazando se tiene que:

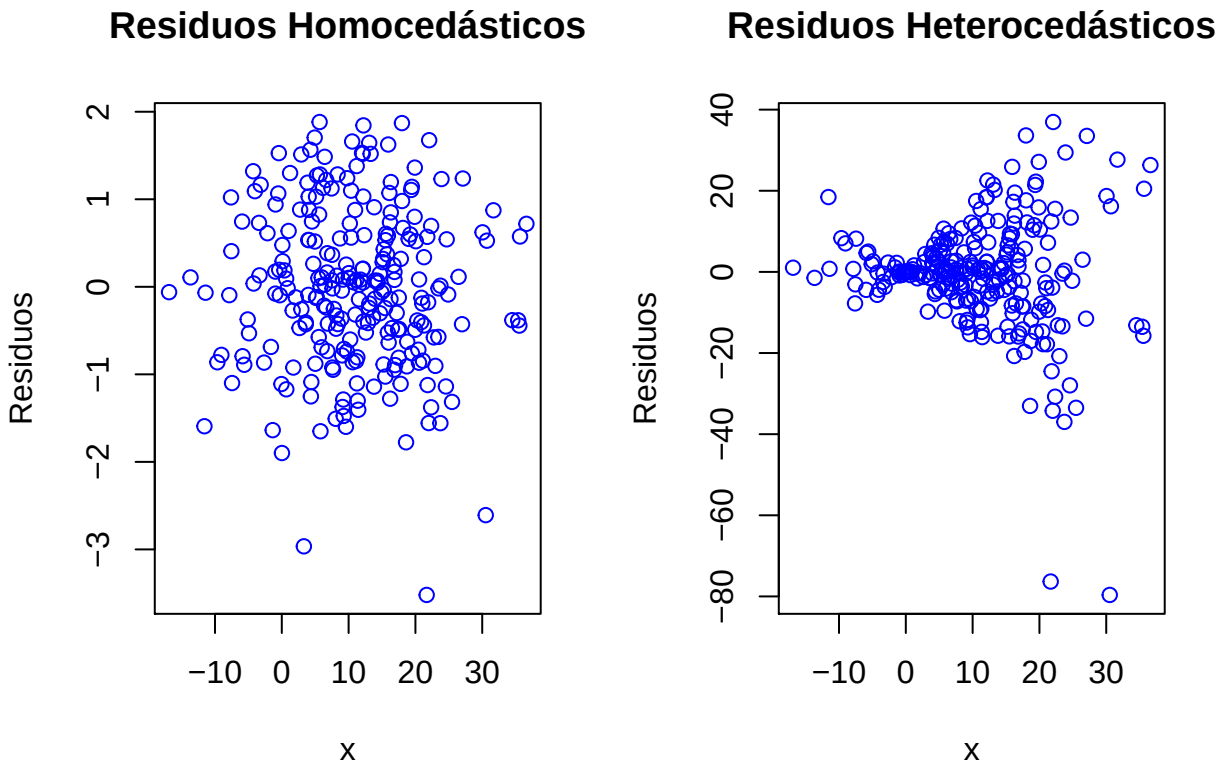
$$var(b_2|x_i) = \frac{\sum_i (x_i - \bar{x})^2 E(\epsilon_i^2|x_i)}{\left(\sum_i (x_i - \bar{x})^2\right)^2}$$

$$var(b_2|x_i) = \frac{\sum_i (x_i - \bar{x})^2 \sigma_i^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \neq \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Detectando la heterocedasticidad

Método 1: Análisis visual de residuos.

Una forma preliminar de indagar sobre la existencia de heterocedasticidad es a través de la inspección visual de los residuos.



Método 2: Contrastes de Hipótesis

Contraste de Multiplicador de Lagrange o Contraste de Breusch - Pagan

Defínase a la función de varianza de y_i como:

$$var(y_i|x_i) = \sigma_i^2 = E(\epsilon_i^2|x_i) = h(\alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is})$$

Donde z_i puede ser igual o diferente que las x_i.

Dos formas funcionales muy utilizadas para $h(.)$ son:

$$h(\alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is}) = exp(\alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is})$$

$$h(\alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is}) = \alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is}$$

Cuando la varianza no depende de variable alguna, se obtiene el caso de varianza homocedástica:

$$h(\alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is}) = h(\alpha_0)$$

En consecuencia, las hipótesis nula y alternativas para un contraste de heterocedásticidad basado en la función de varianza es:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_s = 0$$

H_1 :No todos los α_j en H_0 son cero.

Para construir el estadistico de contraste, defina:

$$v_i = \epsilon_i^2 - E(\epsilon_i|x_i)$$

De tal manera que

$$\epsilon_i^2 = E(\epsilon_i|x_i) + v_i = \alpha_0 + \alpha_1 z_{1i} + ... + \alpha_s z_{si} + v_i$$

Como ϵ_i^2 es no observable, se sustituye por su estimado, el residuo:

$$e_i^2 = \alpha_0 + \alpha_1 z_{1i} + ... + \alpha_s z_{si} + v_i$$

Estimando por MCO esta ecuación, se evalúa si las variables consideradas explican las variaciones en e_i^2. Utilizando el R^2 para medir la proporción de la varianza de ϵ_i explicada por las z_s, se tiene que si la H_0 es cierta el contraste se distribuye:

$$\chi^2 = N \times R^2 \sim \chi^2_{s-1}$$

Contraste de White

Un problema con el contraste anterior es que presupone conocimiento de las variables en la función de varianza si la hipótesis alternativa es cierta. La propuesta de White es sustituir las z_s con las x_i, sus cuadrados y productos cruzados. Bajo las mismas hipótesis, el contraste de White es hecho bajo la prueba Chi cuadrada utilizando el contraste sobre el estadístico $\chi^2 = N \times R^2$.

Corrigiendo la heterocedásticidad.

Como se observó anteriormente, la heterocedásticidad afecta la inferencia a través de su impacto sobre los errores estándar del los estimadores de MCO. La solución por lo tanto pasa por corregir dichos errores estándar para poder realizar la inferencia. Existen dos situaciones bajo las cuales dicha corrección puede ser implimentada:

a) la forma de la heterocedásticidad es conocida.

b) la forma de la heterocedásticidad es desconocida.

En el primero de los casos el estimador resultante es conocido como Mínimos Cuadrados Generalizados o MCG, mientras que en el segundo como Mínimos Cuadrados Generalizados Factibles.

Estimador de MCG

Cuando la forma de la heterocedásticidad es conocida la matriz de covarianzas de ϵ, $E(\epsilon\epsilon'|X) = \Omega_\epsilon$, es conocida. En ese caso, se transforma el modelo premultiplicando por $\Omega_\epsilon^{-1/2}$

$$\Omega_\epsilon^{-1/2} y = \Omega_\epsilon^{-1/2} X\beta + \Omega_\epsilon^{-1/2}\epsilon$$

Definiendo: $y^* = \Omega_\epsilon^{-1/2} y$; $X^* = \Omega_\epsilon^{-1/2} X$ y $\epsilon^* = \Omega_\epsilon^{-1/2}\epsilon$, el MRL es:

$$y^* = X^*\beta + \epsilon^*$$

El estimador de β es:

$$b_{MCG} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (X'\Omega_\epsilon^{-1/2\prime}\Omega_\epsilon^{-1/2}X)^{-1}X'\Omega_\epsilon^{-1/2\prime}\Omega_\epsilon^{-1/2}y$$

Dado que $\Omega_\epsilon^{1/2}$ es una matriz cuadrada y simétrica, se tiene que $\Omega_\epsilon^{-1/2}\Omega_\epsilon^{-1/2} = \Omega_\epsilon$ y que $\Omega_\epsilon^{-1/2'} = \Omega_\epsilon^{-1/2}$, se tiene que:

$$b_{MCG} = (X'\Omega_\epsilon^{-1}X)^{-1}X\Omega_\epsilon^{-1}y$$

Note que en este modelo transformado:

$E(\epsilon^*\epsilon^{*\prime}|X) = E(\Omega_\epsilon^{-1/2}\epsilon\epsilon'\Omega_\epsilon^{-1/2'}|X) = \Omega_\epsilon^{-1/2}E(\epsilon\epsilon'|X)\Omega_\epsilon^{-1/2'} = \Omega_\epsilon^{-1/2}\Omega_\epsilon\Omega_\epsilon^{-1/2'} = I$

Es homocedástica.

Para fines ilustrativos, considere el siguiente modelo de regresión simple con heterocedásticidad:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

donde $E(\epsilon_i|x_i) = 0$, $var(\epsilon_i|x_i) = \sigma_i^2$, $E(\epsilon_i\epsilon_j|x_i) = 0$ para $(i \neq j)$.

Donde se asume que:

$$var(\epsilon_i|x_i) = \sigma_i^2 = \sigma^2 x_i$$

Como la varianza es conocida, se puede realizar una transformación del modelo:

$$\frac{y_i}{\sqrt{x_i}} = \beta_0\frac{1}{\sqrt{x_i}} + \beta_1\frac{x_i}{\sqrt{x_i}} + \frac{\epsilon_i}{\sqrt{x_i}}$$

$$y_i^* = \beta_0 x_{1i}^* + \beta_1 x_{i2}^* + \epsilon_i^*$$

Para mostrar que $\frac{1}{\sqrt{x_i}}$ es la transformación adecuada, note que:

$$var(\epsilon_i^*|x_i) = E(\epsilon_i^{*2}|x_i) = E\left[\left(\frac{1}{\sqrt{x_i}}\epsilon_i\right)^2\middle| x_i\right] = \frac{1}{x_i}E(\epsilon_i^2|x_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2$$

Es decir, es homocedástica.

Estimador de MCGF

En este caso se desconoce la forma exacta de la forma funcional de la varianza heterocedástica. Se parte asumiendo que la varianza depende de una o un conjunto de variables que explican el fenómeno.

Se requiere una estimación de la matriz de covarianzas, $\hat{\Omega}_\epsilon$

Para ilustrar este estimador, cosidere el caso del MRL simple. En este caso, asumiendo que la varianza depende de una o un conjunto de variables que explican el fenómeno, por ejemplo:

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_i)$$

Donde se utiliza la función exponencial para garantizar que la varianza es siempre positiva. Aplicando logatirmo natural:

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i$$

Para la estimación de α_1 y α_2, se utiliza e_i^2 y se reescribe:

$$\ln(e_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_i + v_i$$

La cual se estima por MCO, para obtener:

$$\hat{\sigma}_i^2 = exp(a_1 + a_2 z_i)$$

donde a_1 y a_2 son los valores estimados de α_1 y α_2.

Dividiendo el modelo por la varianza estimada:

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \epsilon_i^*$$

donde $y_i^* = \frac{y_i}{\hat{\sigma}_i}$; x_{i1}^*; $x_i^* = \frac{x_i}{\hat{\sigma}_i}$.

Error de Especificación en el MRL

Contenido

- Error de Especificación en la Forma Funcional
- Variables Proxy variables explicativas no observadas
- Modelos de pendientes aleatorias
- Propiedades MCO con Errores de Medición
- Datos Faltantes y Muestras no Aleatorias

Introducción

Un supuesto de identificación del MRL es:

$$E(\varepsilon_i | x_i) = 0$$

Es decir, que las x_i son variables exógenas. Cuando este supuesto no se cumple, se dice que x_i es una variable explicativa endógena. La implicancia es que los estimadores de MCO están segados:

$$E[b|x] = \beta + E\left[\frac{\sum(x_i - \bar{x})\varepsilon_i}{(x_i - \bar{x})^2} | x_i\right] \neq \beta$$

Hay varias situaciones que pueden generar este resultado:

- Omisión de una variable relevante del modelo.
- Error de medición de una variable explicativa.
- Diseóo de muestras endógenas.
- Forma funcional incorrecta.

Error de Especificación en la Forma Funcional

En lo que sigue nos concentramos en el problema generado por forma funcional incorrecta. Suponga que el modelo verdadero es:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + beta_4 educ \times female + u$$

El error de especificación en este caso podría darse porque:

a. se omite la forma funcional correcta en que debe introducirse experiencia: $exper^2$

b. se omite el término de interacción: $educ \times female$

c. se usa $wage$ en lugar $log(wage)$

Prueba RESET para especificación incorrecta de formas funcionales

RESET: Prueba de Error de Especificaci?n de la Regresión (RESET).

Si el modelo correcto es:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

entonces, al añadir formas funcionales de las variables explicativas ninguna será significativa.

La prueba se basa en estimar:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error$$

Con $H_o : \delta_1 = 0, \delta_2 = 0$, es decir, la forma funcional del modelo es la correcta.

Se usa una prueba F de restricciones lineales con $n - k - 3$ grados de libertad.

Ejemplo: Ecuación para el precio de las viviendas

- Modelo 1:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

- Modelo 2:

$$log(price) = \beta_0 + \beta_1 log(lotsize) + \beta_2 log(sqrft) + \beta_3 bdrms + u$$

- donde: $price$: Precio de la vivienda, \$1000s $lotsize$: tamaño del lote (solar) en pies cuadrados $sqrft$: tamaño de la casa en pies cuadrados $bdrms$: número de habitaciones

Ejemplo: Ecuación para el precio de las viviendas

```
library(lmtest)
library(stargazer)
library(wooldridge)
data("hprice1")
attach(hprice1)
#Modelo lin-lin
mod.lin <-lm(price~lotsize+sqrft+bdrms)
#Modelo log-log
mod.log <-lm(lprice~llotsize+lsqrft+bdrms)
```

Table 3: Resultado Estimaciones

	Dependent variable:	
	price	lprice
	(1)	(2)
lotsize	0.002^{***}	
	(0.001)	
sqrft	0.123^{***}	
	(0.013)	
llotsize		0.168^{***}
		(0.038)
lsqrft		0.700^{***}
		(0.093)
bdrms	13.853	0.037
	(9.010)	(0.028)
Constant	-21.770	-1.297^{**}
	(29.475)	(0.651)
Observations	88	88
R^2	0.672	0.643
Adjusted R^2	0.661	0.630
Residual Std. Error (df = 84)	59.833	0.185
F Statistic (df = 3; 84)	57.460^{***}	50.424^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Ejemplo: Ecuación para el precio de las viviendas

```
library(lmtest)

#Modelo lin-lin
resettest(mod.lin,power=2:3)
```

```
##
##  RESET test
##
## data:  mod.lin
## RESET = 4.6682, df1 = 2, df2 = 82, p-value = 0.01202
```

```
#Modelo log - log
resettest(mod.log,power=2:3)
```

```
##
##  RESET test
##
## data:  mod.log
## RESET = 2.565, df1 = 2, df2 = 82, p-value = 0.08308
```

Observaciones sobre la Prueba RESET

- No provee indicación sobre cómo proceder si se rechaza el modelo.
- No se puede usar para problemas de especificación relativos a omisi?n de variables no observadas.
- Es solo una prueba para forma funcionales.

Variables proxy para variables explicativas no observadas

- Hay situaciones en las que no se observa una o un grupo de las variables explicativas.
- Estimar el modelo con las observables deriva en estimadores sesgados, si alguna de estas variables está correlacionada con la omitida.
- El uso de variables proxy de la omitida puede resolver o atenuar el problema.

Variables proxy para variables explicativas no observadas

- Sea el modelo verdadero:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

-Suponga que no se observa x_3^*, pero se tiene una proxy x_3 tal que:

$$x_3^* = \delta_0 + \delta_1 x_3 + v_3$$

La solución al problema de variables omitidas consiste en usar x_3 en el modelo en lugar de x_3^*. Para que esta estrategia funcione, es decir, proporcione estimadores consistentes de β_1 y β_2, tienen que cumplirse los siguientes supuestos:

1. u no está correlacionado con x_1, x_2 y x_3^*.
2. u no está correlacionada con x_3. Es decir, incluidas x_1,x_2 yx_3^* en el modelo x_3 es irrelvante.
3. v_3 no está correlacionado con x_1, x_2 y x_3.

El supuesto 3 es:

$$E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) = \delta_0 + \delta_1 x_3$$

Es decir, x_3^* tiene correlaci?n cero con x_1 y x_2 una vez descontados los efectos parciales de x_3.

Estos supuestos son suficientes para que esta estrategia funcione. Sustituyendo:

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3$$

Defínase el error compuesto:

$$e = u + \delta_3 v_3$$

Dado $E(u|x_1, x_2, x_3) = E(v_3|x_1, x_2, x_3) = 0$ entonces $E(e|x_1, x_2, x_3) = 0$.

Así el modelo es:

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

Por lo que obtenemos estimados insesgados de β_1 y β_2. As? como de α_0 y α_3.

Table 4: Resultado Estimaciones

	Dependent variable:	
	lsalario	
	(1)	(2)
educ	0.065^{***}	0.054^{***}
	(0.006)	(0.007)
exper	0.014^{***}	0.014^{***}
	(0.003)	(0.003)
antig	0.012^{***}	0.011^{***}
	(0.002)	(0.002)
casado	0.199^{***}	0.200^{***}
	(0.039)	(0.039)
sur	-0.091^{***}	-0.080^{***}
	(0.026)	(0.026)
urbano	0.184^{***}	0.182^{***}
	(0.027)	(0.027)
afro	-0.188^{***}	-0.143^{***}
	(0.038)	(0.039)
IQ		0.004^{***}
		(0.001)
Constant	5.395^{***}	5.176^{***}
	(0.113)	(0.128)
Observations	935	935
R^2	0.253	0.263
Adjusted R^2	0.247	0.256
Residual Std. Error	0.365 (df = 927)	0.363 (df = 926)
F Statistic	44.747^{***} (df = 7; 927)	41.265^{***} (df = 8; 926)
Note:		*p<0.1; **p<0.05; ***p<0.01

Modelos con pendientes aleatorias

Hasta ahora el supuesto ha sido de que los coeficientes de las pendientes es el mismo para todos los individuos. O si varían lo hace en características medibles: sexo, estado civil, etc. Una especificaciín alternativa es el modelo de coeficiente aleatorio:

$$y_i = a_i + b_i x_i$$

donde b_i se considera como una observación muestreada aleatoriamente de la población. Donde el modelo visto supone que $a_i = u_i$ y $b_i = \beta$. Con una muestra de N datos, es imposible estimar este modelo, sino esperar estimar la pendiente e intercepto promedios.

$$\alpha = E(a_i)$$

y $\beta = E(b_i)$.

Es decir, β es el promedio del efecto parcial de x sobre y, es decir el *efecto parcial promedio (EPP)*. Escribiendo

$$a_i = \alpha + c_i$$
$$b_i = \beta + d_i$$

Por construcción $E(d_i) = E(c_i) = 0$. Sustituyendo

$$y_i = \alpha + \beta x_i + c_i + d_i x_i = \alpha + \beta x_i + u_i$$

donde $u_i = c_i + d_i x_i$

Dados los supuestos sobre c_i y d_i MCO producirá estimadores insesgados. No obstante, el error contendrá heterocedásticidad

$$var(u_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$$

A menos que $\sigma_d^2 = 0$, que en ese caso $b_i = \beta$ para toda i.

Propiedades de MCO bajo error de medición

Existe la posibilidad que la variable económica de interés pueda ser medida con error. Este error de medición estará contenido en el modelo. A diferencia del caso de variable omitida, el interés puede ser sobre el coeficiente de la variable que contiene el error.

Tipos de errores de medición:

1. en las variables dependientes
2. en las variables explicativas

Considere el siguiente modelo

$$y^* = \beta_0 + \beta_1 x_1 + u$$

donde y^* es la variable de interés, se asume que hay una variable y con un ligero error de medidción, definido como:

$$e_0 = y - y^*$$

El modelo estimable es:

$$y = \beta_0 + \beta_1 x_1 + u + e_0$$

Si e_0 no está correlacionado con x_i, MCO es insesgado y consistente.

Lo que si es cierto, bajo esos supuestos

$$var(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2.$$

Error de Medición en las Variables Explicativas

Suponga el siguiente modelo:

$$y = \beta_0 + \beta_1 x^* + u$$

donde se cumplen los supuestos de Gauss Markov, pero x^* es una variable no observda, con una medición con error.

$$e_1 = x_1 - x_1^*$$

.

Se supone que en la poblacion $E(e_1) = 0$.

Las propiedades de MCO dependen de los supuestos que se hagan sobre el error una vez se sustituya la variable del modelo por la variable medida con error:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

este error compuesto, tiene media cero, por qué?

La varianza es:

$$var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_e^2$$

Es decir, el error de medición hace que aumente la varianza del error. El supuesto de errores cl?sicos en las variables (ECV), dice que:

$$cov(x_1^*, e) = 0$$

La medición observada es:

$$x_1 = x_1^* + e_1$$

Asimismo,

$$cov(x_1 e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

Respecto al modelo con el error compuesto:

$$cov(x_1, u - \beta_1 e_1) = -\beta cov(x_1 e_1) = -\beta \sigma_{e_1}^2$$

Por tanto, en el caso de ECV, al regresi?n de MCO ofrece estrimados sesgados e inconsistentes.

En particular,

$$plim(b_1) = \beta_1 + \frac{cov(x_1, u - \beta_1 e)}{var(x_1)}$$

$$= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{var(x_1)}$$

$$= \beta_1 (1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2})$$

$$= \beta_1 (\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}) < \beta_1$$

A esto se le llama sesgo de atenuación.

Datos Faltantes

En ocasiones los datos que disponemos padecen de datos faltantes en algunas de las variables clave del modelo. El efecto estadístico sobre la estimación depende de la razón de los datos faltantes:

1. Datos faltantes al azar: la estimación es insesgada, aunque menos eficiente.
2. Datos faltantes por muestreo no aleatorio

Muestras no aletorias

En este caso significa que algunas observaciones tienen menos probabilidad de ser muestreadas.

Tipos de selección muestral:

1. selección muestral exógena: elección de la muestra en base a las variables exógenas. No causa sesgo en el estimador de MCO.
2. selección muestral endógena: elección muestral basada en la variable dependiente. Causa sesgo en MCO.

Regresores endógenos y Variables Instrumentales

Introducción

En el modelo de regresión lineal (MRL):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

El supuesto de exogeneidad de las explicativas implica que

$$cov(x_i, u) = 0$$

para todo $i = 1, ..., k$

Se dice que un regresor es endógeno cuando este supuesto es violado. Es decir,

$$\frac{\partial E[y|X]}{\partial x_j} = \beta_j + \frac{\partial E[u|X]}{\partial x_j}$$

¿Cómo surge la endogeneidad?

Ejemplo 1: Sesgo por variables omitidas

Suponga que el verdadero modelo de demanda de gasolina es:

$$ln(G) = \beta_1 + \beta_2 ln(Precio) + \beta_3 ln(Ingreso) + u$$

Suponga que no se incluye el ingreso y se estima:

$$ln(G) = \beta_1 + \beta_2 ln(Precio) + w$$

donde $w = \beta_2 ln(Ingreso) + u$

Si Ingreso está relacionada con Precio, β_1 y β_2 no se pueden estimar de manera consistente.

Este sesgo se denomina **Sesgo por variable omitida**

Ejemplo 2: Efecto tratamiento endógeno

Suponga que desea analizar el efecto de programas de becas sobre la probabilidad de encontrar empleo. Sea U una dummy igual a 1 si tuvo beca y 0 si no. El modelo a estimar es:

$$lnY = \beta X + \delta U + v$$

La estimación por MCO genera una estimación sesgada, debido a su correlación con factores no observables presentes en v.

Ejemplo 3: Ecuaciones Simultáneas

Considere el siguiente modelo de oferta y demanda de un producto:

$$D = \alpha_0 + \alpha_1 Precio + \alpha_2 Ingreso + \varepsilon_d$$

$$O = \beta_0 + \beta_1 Precio + \beta_2 Costos + \varepsilon_o$$

$$O = D$$

Considere una estimación de los parámetros de la función de demanda usando los valores observados de precios y cantidades de equilibrio.

Usando las condiciones de equilibrio y despejando para el precio, tenemos:

$$Precio = (\alpha_0 - \beta_0 + \alpha_2 Ingreso - \beta_2 Costo + \varepsilon_d - \varepsilon_o)/(\beta_1 - \alpha_1)$$

Note que Precio está correlacionado con ε_d y ε_o, por lo tanto la estimación por MCO de los parámetros de la demanda están sesgados.

Este es el sesgo de **Ecuaciones Simultáneas o de Simultaneidad**

Ejemplo 4: Errores de Medición

El modelo básico para analizar los efectos de la eduación sobre el ingreso de los individuos es la "ecuación de Mincer (1974)":

$$y = \beta_1 + \beta_2 Educacion + \varepsilon$$

La variable educación es no observable, y se aproxima usando los años de escolaridad.

$$Escolaridad = Educacion + u$$

Por lo tanto el modelo a estimar es:

$$y = \beta_1 + \beta_2 Escolaridad + \varepsilon$$

donde $w = \varepsilon - \beta_2 u$. Por lo que escolaridad está correlacionada con el error

Este es el **Sesgo de Atenuación**.

Ejemplo 5: Muestreo no aleatorio

Suponga que se desea estimar la relación entre la riqueza y otros factores:

$$riqueza = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u$$

Suponga que solo se incluyen personas cuya riqueza sea menor a RD$1 millón de pesos.

Esta es una muestra no aleatoria de la población interés.

Este es el **Sesgo de Selección Muestral**

Ejemplo 6: Abandono o "desgaste" (attrition)

Suponga que desea analizar el impacto, en el tiempo, de un programa de entrenamiento sobre la productividad de un conjunto de empresas.

Al final del periodo, las observaciones existentes son un subconjunto de las de inicio del periodo, las m?s productivas.

Por lo que no son una muestra representativa de la población. A este se le conoce como **Sesgo de Supervivencia**

Ejemplo 7: Abandono o "desgaste" (attrition)

Suponga un análisis de los efectos de un medicamento para la colitis. En la medida que algunos individuos de la muestra muestren mejoria pueden dejar de asistir a los tratamientos dejando la muestra a un grupo no representativo de la población. A este se le conoce como **Sesgo por Abandono, Desgaste o 'Attrition'**

Enfoques para solucionar el problema de la endogeneidad.

- Especificación Estructural
- Variables Instrumentales

El Estimador de Variables Instrumentales

El método de variables instrumentales consiste en identificar una variable z que satisfaga dos supuestos:

1. Exogeneidad: $Cov(z, u) = 0$
2. Relevancia: $Cov(z, x) \neq 0$

Utilizar esta información para estimar de manera insesgada y consistente los parámetros del modelo.

El supuesto de exogeneidad se sustenta en base al comportamiento econ?mico o introspección.

El supuesto de relevancia se puede sustentar a través de:

$$x = \pi_0 + \pi_1 z + v$$

Entonces, dado que $\pi_1 = Cov(z, x)/Var(z)$, el supuesto de relevancia se mantiene si y solo si $\pi_1 \neq 0$. Es decir, rechazar

$$H_0 : \pi_1 = 0$$

versus

$$H_1 : \pi_1 \neq 0$$

Ejemplo: Retornos de la educación

Considere la siguiente ecuación de salarios:

$$log(salario) = \beta_0 + \beta_1 educ + \beta_2 habilidad + e$$

Suponga que la variable *habilidad* no es observable, por lo que el modelo a estimar es:

$$log(salario) = \beta_0 + \beta_1 educ + u$$

Si *educ* y *habilidad* están correlacionadas, el estimador de MCO de β_1 está sesgado. ?Por qué?

El uso del método de Variables Instrumentales (VI) requiere disponer de un instrumento para *educ* que:

1. no está correlacionada con *habilidad* y con ning?n otro factor que afecte el salario.
2. está correlacionada con *educ*

Propuestas

1. z = último número de la cédula.

- Cumple el supuesto de exogeneidad. ¿Por qué?
- No cumple el supuesto de relevancia. ¿Por qué?

2. z = IQ (variable proxy)

- Cumple el supuesto de relevancia. ¿Por qué?
- No cumple el supuesto de exogeneidad. ¿Por qué?

3. z = educación de la madre
4. z = número de hermanos.

El estimador de VI

VI permite estimar de manera consistente los parámetros del modelo. A este proceso se le denomina **identificación**. El punto de partida es la covarianza poblacional entre z e y:

$$Cov(z, y) = E[(z - \bar{z})(y - \bar{y})]$$

$$= E[(z - \bar{z})(\beta_1(x - \bar{x}) + u)]$$

$$= E[\beta_1(z - \bar{z})(x - \bar{x})] + E[(z - \bar{z})u]$$
$$= \beta_1 Cov(z, x) + Cov(z, u)$$

Dados los supuestos $Cov(z, x) \neq 0$ y $Cov(z, u) = 0$,

$$\beta_1 = \frac{Cov(z, y)}{Cov(x, z)}$$

Sustiyendo por los valores muestrales:

$$b_1^{VI} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y - \bar{y})}{\sum_{i=1}^{N}(z_i - \bar{z})(x - \bar{x})}$$

Este es el estimador de VI. MCO es un caso especial de VI cuando $z = x$.

Al igual que MCO, el estimador de VI tiene una distribución normal en muestras grandes.

Para computar el error estándar del estimador se supone homocedasticidad condicionando en z

$$E(u|z) = \sigma^2$$

Sin embargo la varianza asint?tica de b_1^{VI} es:

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

donde σ_x^2 es la varianza poblacional de x, y $\rho_{x,z}$ es el cuadrado de la correlación poblacional entre x y z.

Esta expresión nos permite obtener un error estándar para el b_1^{VI}. La versión muestral es:

$$\frac{\hat{\sigma}^2}{STC_x R_{x,z}^2}$$

Asimismo, nos permite comparar los errores estándar de los estimados de MCO y VI. Estas solo difieren por $R_{x,z}^2$ y como este está entre 0 y 1, entonces, cuando MCO es válido:

$$ee(\hat{b_1}^{VI}) > ee(\hat{b_1}^{MCO})$$

- Cuanto mayor es la relación entre x y z, mayor $R_{x,z}^2$, menor la varianza del estimador de VI.

Table 5: Resultado Estimaciones

	Dependent variable:		
	lsalario	educ	lsalario
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
educ	0.109^{***}		0.059^{*}
	(0.014)		(0.035)
educ_padre		0.282^{***}	
		(0.021)	
Constant	-0.185	9.799^{***}	0.441
	(0.185)	(0.199)	(0.446)
Observations	428	753	428
R^2	0.118	0.196	0.093
Adjusted R^2	0.116	0.195	0.091
Residual Std. Error	0.680 (df = 426)	2.046 (df = 751)	0.689 (df = 426)
F Statistic	56.929^{***} (df = 1; 426)	182.812^{***} (df = 1; 751)	
Note:			*p<0.1; **p<0.05; ***p<0.01

Propiedades de VI con una variable instrumental deficiente.

- La correlación débil entre z y x puede tener consecuencias serias:

$$plim\ b_1^{VI} = \frac{Cov(z,y)}{Cov(z,x)} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \times \frac{\sigma_u}{\sigma_x}$$

- si $Corr(z,x)$ es pequeña la inconsistencia del estimador de VI puede ser muy grande.
- A este tipo de problema pertenece el caso de los **instrumentos d?biles**.

Estimación de VI del MRLM

- Considere el siguiente MRLM

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

- Se denomina **ecuación estructural**, enfatizando que el interés es sobre las β_j de la relación causal.
- Suponga que y_2 es la variable explicativa endógena, mientras que z_1 es exógena.
- Se necesita un instrumento z_2, debido a que z_1 ya está incorporada en la ecuación, tal que

$$E(u_1) = 0;\ Cov(z_1,u) = 0;\ Cov(z_2,u) = 0$$

- desarrollando los momentos muestrales

$$\sum_{i=1}^{N}(y_{i1} - b_0 - b_1 y_{i2} - b_2 z_{i1}) = 0$$

$$\sum_{i=1}^{N} z_{i,1}(y_{i1} - b_0 - b_1 y_{i2} - b_2 z_{i1}) = 0$$

$$\sum_{i=1}^{N} z_{i2}(y_{i1} - b_0 - b_1 y_{i2} - b_2 z_{i1}) = 0$$

Los estimadores resultantes son los de VI. Cuando se escribe la variable explicativa endógena, en t?rminos de las exógenas, se denomina **forma reducida**

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$$

Mnimos Cuadrados en Dos Etapas

¿Qué sucede cuando tenemos más instrumentos que variables explicativas exógenas?

Suponga que en el modelo anterior tengamos dos instrumentos para la variable explicativa endógena.

Se conoce como *restricciones de exclusión* al supuesto de que los instrumentos no aparencen en la ecuación estructural.

En el caso de más de un instrumento por regresor endógeno, se estima la ecuación de forma reducida:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

La mejor VI para y_2 es el valor ajustado de la regresión anterior.

A continuación se procede a estimar el modelo con los valores ajustados de y2 como regresor.

A este procedimiento se le denomina MC2E.

Table 6: Resultado Estimaciones

	Dependent variable:		
	lsalario	educ	lsalario
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
educ	0.107^{***}		0.061^{*}
	(0.014)		(0.031)
exper	0.042^{***}	0.085^{***}	0.044^{***}
	(0.013)	(0.026)	(0.013)
exper2	-0.001^{**}	-0.002^{**}	-0.001^{**}
	(0.0004)	(0.001)	(0.0004)
educ_padre		0.185^{***}	
		(0.024)	
educ_madre		0.186^{***}	
		(0.026)	
Constant	-0.522^{***}	8.367^{***}	0.048
	(0.199)	(0.267)	(0.400)
Observations	428	753	428
R^2	0.157	0.262	0.136
Adjusted R^2	0.151	0.258	0.130
Residual Std. Error	0.666 (df = 424)	1.964 (df = 748)	0.675 (df = 424)
F Statistic	26.286^{***} (df = 3; 424)	66.520^{***} (df = 4; 748)	
Note:			*p<0.1; **p<0.05; ***p<0.01

Multicolinealidad y MC2E

En el caso de MC2E la multicolinealidad puede ser un problema potencial.

Esto es debido a la relación de colinealidad que la variable explicativa endógena con las demás variables explicativas exógenas.

En el MRLM, la varianza de β_1 se aproxima con:

$$\frac{\sigma^2}{S\hat{T}C_2(1-\hat{R}_2^2)}$$

donde $\hat{R}_2^2$ es la R cuadrada de una regresión de $\hat{y}_2$ sobre todas las demás exógenas que aparecen en la ecuación estructural.

Múltiples variables explicativas endógenas

Considere el siguiente MRLM

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1$$

Suponga que y_2 y y_3 son variables explicativas endógenas.

Para estimar este modelo con MC2E, se necesitan *al menos* dos instrumentos: dos variables exógenas que no aparezcan en la ecuación estructural.

A esto se le denomina **Condición de Orden**: es solo una condición necesaria.

Hay otra condición conocida como **Condición de Rango**, la cual es una condición suficiente.

Pruebas de Endogenidad

Al igual que los otros casos de violación de supuestos, una pregunta es ¿Cómo detecto la presencia de endogeneidad en mi modelo?

Muchas veces se puede inferir este problema mediante razonamiento econ?mico.

Sin embargo, Hausman (1978) provee un contraste de endogeneidad.

Prueba de Endogeneidad de Hausman

1. Estime la forma reducida de la variable explicativa endógena mediante su regresión sobre todas las exógenas de la ecuación estructural y los instrumentos. Guarde los residuos($\hat{v}$).
2. Agregue $\hat{v}$ en la ecuación estructural y evalue si el coeficiente de $\hat{v}$ es estadísticamente diferente de cero. Si es así, entonces se concluye que hay endogeneidad.

La H_0 es que no hay endogeneidad.

Utilice una prueba t robusta a la heterocedásticidad.

Table 7: Resultado Estimaciones

	Dependent variable:		
	lsalario	educ	lsalario
	(1)	(2)	(3)
educ	0.107^{***}		0.064^{**}
	(0.014)		(0.029)
exper	0.042^{***}	0.085^{***}	0.046^{***}
	(0.013)	(0.026)	(0.013)
exper2	-0.001^{**}	-0.002^{**}	-0.001^{**}
	(0.0004)	(0.001)	(0.0004)
educ_padre		0.185^{***}	
		(0.024)	
educ_madre		0.186^{***}	
		(0.026)	
v			0.056^{*}
			(0.033)
Constant	-0.522^{***}	8.367^{***}	-0.011
	(0.199)	(0.267)	(0.359)
Observations	428	753	428
R^2	0.157	0.262	0.163
Adjusted R^2	0.151	0.258	0.155
Residual Std. Error	0.666 (df = 424)	1.964 (df = 748)	0.665 (df = 423)
F Statistic	26.286^{***} (df = 3; 424)	66.520^{***} (df = 4; 748)	20.529^{***} (df = 4; 423)
Note:			*p<0.1; **p<0.05; ***p<0.01

Cuando existe más de un instrumento para la variable explicativa endógena, se corre el riesgo que la combinación lineal de estos generada en la primera etapa de MC2E se correlacione con el error del modelo estructural, ya sea por:

1. Uno o más de las candidatas a instrumentos está correlacionada con el error.
2. Los residuales de la ecuación estructural se relaciona con la combinaciones lineales de los instrumentos.

Prueba de restricciones de sobreidentificación

La **Prueba de Restricciones de Sobreidentificación** compara las estimaciones de VI para un mismo parámetro.

1. Se estima la ecuación estructural mediante MC2E y se obtienen los residuales, $\hat{u}_1$.
2. Se realiza una regresión de $\hat{u}_1$, sobre todas las variables exógenas. Se guarda la R cuadrada: $\hat{R}_1^2$.
3. Con base a la H_0 de que todas las VI no están correlacionadas con u_1, $n\hat{R}_1^2$ se distribuye como χ_q^2, donde q es el número de VI instrumentales al modelo menos el número de variables explicativas endógenas.

Si $n\hat{R}_1^2$ excede el valor crítico de 5% en la distribución χ_q^2, se rechaza la H_0 y se dice que las VI no son exógenas.

Modelos para variables Dependientes Cualitativas y Limitadas

Introducción

En el modelo de regresión lineal:

$$y = X\beta + \epsilon$$

La variable y se suponía una variable continua con distribución

$$y \sim iid(X\beta, \sigma^2)$$

Esta característica cuantitativa permitía:

- Deducir efectos marginales y parciales.

- Realizar predicciones y construir intervalos de confianza.

Table 8: Resultado Estimaciones

	Dependent variable:	
	lsalario	u
	(1)	(2)
educ	0.107^{***}	
	(0.014)	
exper	0.042^{***}	0.001
	(0.013)	(0.013)
exper2	-0.001^{**}	−0.00004
	(0.0004)	(0.0004)
educ_padre		−0.003
		(0.011)
educ_madre		−0.014
		(0.012)
Constant	-0.522^{***}	0.162
	(0.199)	(0.139)
Observations	428	428
R^2	0.157	0.006
Adjusted R^2	0.151	−0.003
Residual Std. Error	0.666 (df = 424)	0.665 (df = 423)
F Statistic	26.286^{***} (df = 3; 424)	0.649 (df = 4; 423)
Note:		*p<0.1; **p<0.05; ***p<0.01

Habíamos considerado la posibilidad de considerar variables discretas o de origen cualititativo como explicativas.

$$y = X\beta + \delta D + \epsilon$$

Donde

$$D = \begin{cases} 1 \\ 0 \end{cases}$$

Y el resultado era que no cambiaba el método de estimación y la interpretación del coeficiente era directa.

Ahora veremos el caso donde el objetivo es considerar modelos donde la variable dependiente es discreta o presenta una discontinuidad en algún tramo de su distribución.

En el caso discreto, estamos ante **información de naturaleza cualitiva**.

En el caso continuo con discontinuidad se dice que es una **variable dependiente limitada**.

Dependiendo el tipo de variable cualitativa o limitada tendremos un tipo o tipos de modelos apropiados para el análisis de sus determinantes y predicción.

Uno de los supuestos que tendremos que abandonar para dar paso a esta nueva familia de modelos, es el de linealidad.

Esto implica modificar el método de estimación por uno que considere la estimación de modelos no lineales.

Variables dependientes cualitativas

Estas variables surgen al solo poder observar un fenómeno a traves de la captura de la elección o respuesta final.

En rigor, el interés es inferir el proceso subyacente, que no es observable.

En ese sentido parten la formulación de lo que se conoce como modelo de elección discreta.

El proceso subyacente se modela a partir de una variable continua y el componente observado, en términos de la probabilidad de que se alcance cierto umbral.

Tipos de variables dependientes cualitativas

- Elección binaria
  - Modelos: MPL, Probit, Logit
- Elección multiple
  - Modelos: Logit y probit multinomiales, probit condicional
- Conteo de eventos
  - Modelos: Poisson

Modelos para variables dependiente binaria

- Muchas elecciones que hacen los indivisuos son del tipo “si o no”, “esa o aquella”.
- Esas elecciones son representadas por un indicador binario.
- Cuando es la variable dependiente, este hecho afecta nuestra elección de modelo estadístico.
- Dando paso a los modelos de **elección binaria**.

Modelo de probabilidad lineal

- Se refiere al MRL cuando la variable dependiente es binaria.

$$y_i = x_i'\beta + \epsilon_i$$

- donde

$$y_i = \begin{cases} 1 \\ 0 \end{cases}$$

- Con $E\{\epsilon_i\} = 0, E\{y_i|x_i\} = x_i'\beta$, entonces,

$$E\{y_i|x_i\} = p\{y_i = 1|x_i\} \times 1 + p\{y_i = 0|x_i\} \times 0 = p\{y_i = 1|x_i\}$$

- La distribución de probabilidad de y y ϵ es:

y	ϵ	Probabilidad
1	$1 - x_i'\beta$	$p = x_i'\beta$
0	$-x_i'\beta$	$1 - p = 1 - x_i'\beta$

- Un problema del MRL es que los errores son heterocedásticos.
- La varianza es (demuestre):

$$var\{\epsilon_i\} = x_i'\beta(1 - x_i'\beta)$$

- Por lo que se puede utilizar MCGF para la estimación eficiente. La varianza estimada es:

$$\hat{\sigma}_i^2 = \hat{var\{\epsilon_i\}} = x_i'b(1 - x_i'b)$$

- Por lo que el modelo puede ser estimado con los datos transformados:

$$y_i^* = x_i^{*\prime}\beta + \epsilon_i^*$$

- Con:

$$y_i^* = y_i/\hat{\sigma}_i$$

$$x_i^* = x_i/\hat{\sigma}_i$$

Debilidades del MPL

- Varias dificultades surgen al implementar este modelo:
  - Las probabilidades pueden ser mayores a uno o negativas.
  - x_i tiene un efecto constante sobre la probabilidad de elegir $y = 1$

$$\frac{\partial p}{\partial x_i} = \beta$$

Modelo de Elección Binaria

- Una solución a las limitaciones del MPL es considerar una función $G(.)$, tal que:

$$p(y = 1|X) = G(X\beta)$$

- donde $0 \leq G(.) \leq 1$

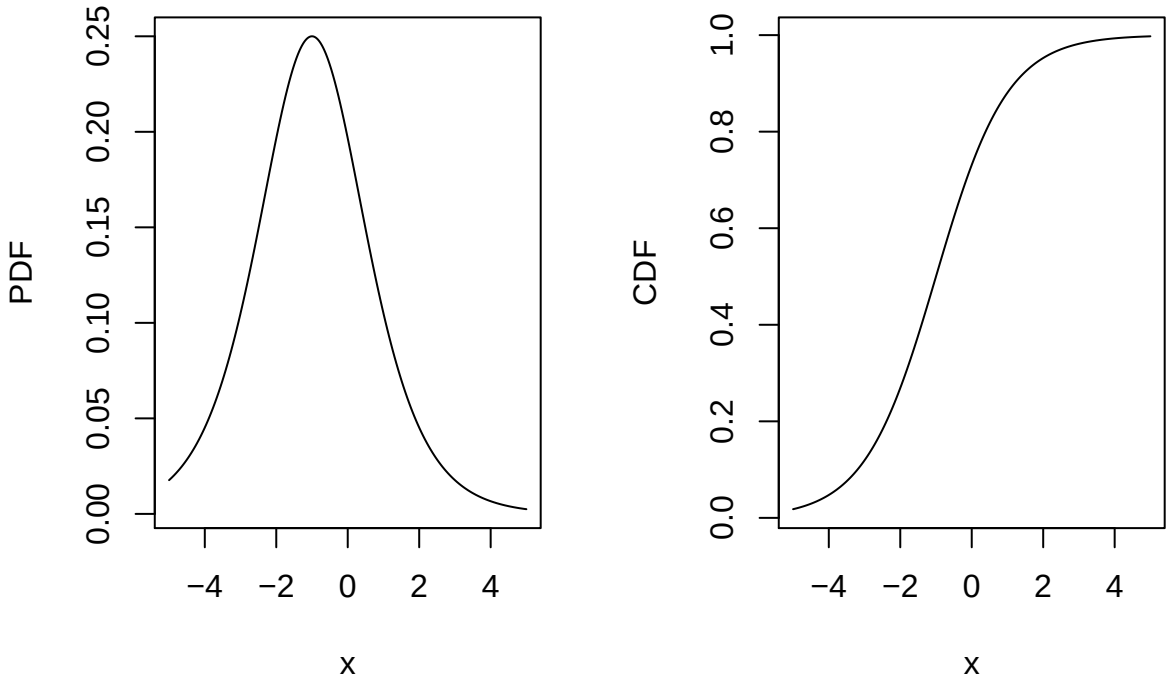
- Dependiendo de la especificación se tendrán distintos modelos.
- El modelo logit surge cuando $G(.)$ es la función logística, $G(z) = \frac{e^z}{1+e^z}$

```
pdf=function(x,mu,s){
  k=(x-mu)/s
  return(exp(-k)/(s*(1+exp(-k))^2))
}

cdf=function(x,mu,s){
  k=(x-mu)/s
  return(1/(1+exp(-k)))
}
x=seq(-5,5,0.01)

## PDF
layout(matrix(1:2,nrow=1))
plot(x,pdf(x,-1,1),type="l",ylab="PDF")

## CDF
plot(x,cdf(x,-1,1),type="l",ylab="CDF")
```

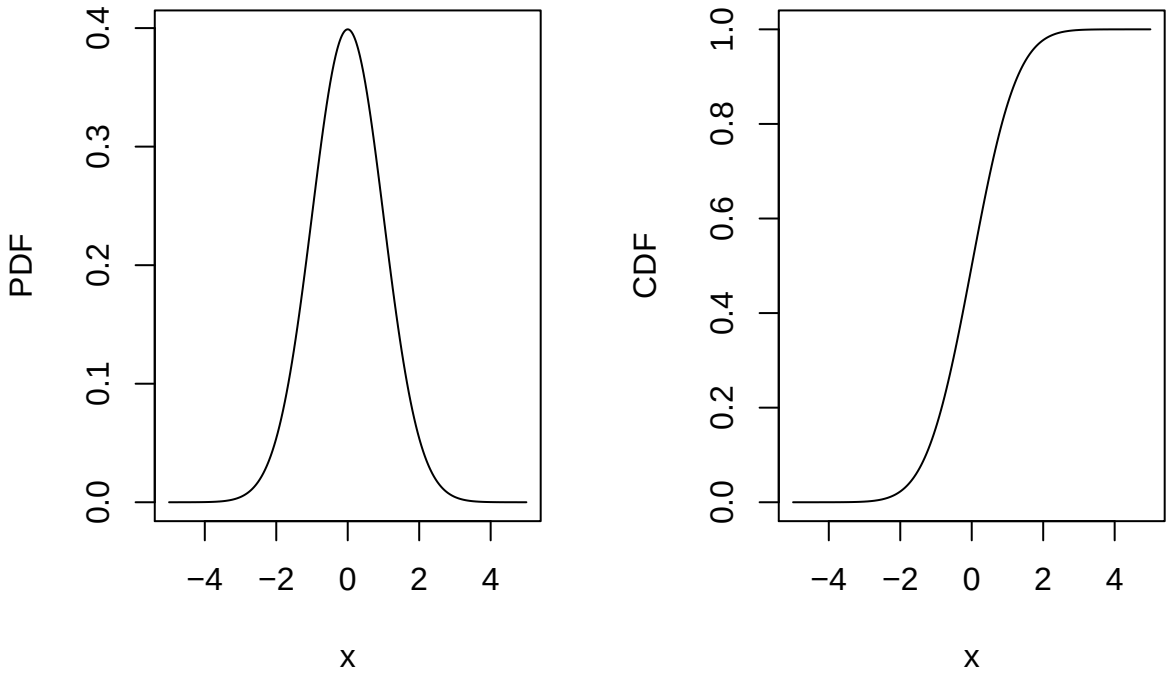


- Cuando $G(.)$ es la distribución de probabilidad acumulada normal, el modelo se denomina probit: $G(.) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv$ y $\phi(z) = (2\pi)^{0.5}e^{-0.5z^2}$

```
x=seq(-5,5,0.01)

## PDF
layout(matrix(1:2,nrow=1))
plot(x,dnorm(x,0,1),type="l",ylab="PDF")

## CDF
plot(x,pnorm(x,0,1),type="l",ylab="CDF")
```



Especificación del modelo de elección binaria

- Supongamos que tenemos el siguiente modelo sugerido por la teoría

$$y^* = X\beta + \epsilon$$

- y^*: variable latente
- Como y^* es no observable sino que se observa su elección exp post, se construye la siguiente función índice.

$$y_i = \begin{cases} 1 & si \quad y^* > 0 \\ 0 & si \quad y^* \leq 0 \end{cases}$$

- Se supone que $E(\epsilon|X) = 0$ y además $\epsilon \sim N(0, \sigma^2)$.
- La probabilidad de respuesta es:

$$p(y = 1|x) = p(y^* > 0|x) = p(X\beta + \epsilon > X) = p(\epsilon > -X\beta|X) =$$

$$1 - G(-X\beta) = G(X\beta)$$

- El objetivo es explicar los efectos de las X sobre la probabilidad de respuesta.

Ejemplo: Determinantes de la fertilidad

- Se quiere estudiar las desiciones de las mujeres para tener hijos en un país en desarrollo, bajo el conjetura de que la reducción de la fertilidad permite que las mujeres trabajen y reduce el número de dependientes en la economía.
- El modelo sugerido viene dado por:

$$y^* = \beta_0 + \beta_1\ edad + \beta_2\ educación + \beta_3\ casada + ...$$
$$\beta_4\ adultos + \beta_5\ educ\ padre + \beta_6\ educ\ madre + \epsilon$$

- Donde y^* es la utilidad o ganancia latente de tener hijos.
- Para estimar, se define $y = 1$ si la mujer ha tenido hijos.
- Por lo tanto, el modelo a estimar es:

$$p(y = 1|X) = G(\beta_0 + \beta_1\ edad + \beta_2\ educación + \beta_3\ casada + ...$$
$$\beta_4\ adultos + \beta_5\ educ\ padre + \beta_6\ educ\ madre)$$

- El mismo sera un logit o un probit según $G(.)$ sea una logística o una normal respectivamente.

Interpretación de los efectos de las variables sobre la probabilidad de respuesta.

- Lo que se busca es el efecto parcial, condicionado por la distribución de probabilidad:

$$p(X\beta) = G(X\beta)$$

$$\frac{\partial p(X\beta)}{\partial x_j} = g(X\beta)\beta_j$$

$$g(z) = \frac{\partial G(z)}{\partial z}$$

- Esto indica que el efecto de x_j sobre $p(X\beta)$ depende del resto de las variables independientes a través de la cantidad positiva $g()$ y el signo del efecto viene dado por β_j.
- Si en el modelo existen variables explicativas binarias:

$$y = X\beta + \delta D + \epsilon$$

- El efecto parcial se computa como:

$$G(X\beta + \delta) - G(X\beta)$$

- Note que el efecto parcial sigue condicionado a los valores de las otras X's.

Estimación

- Como son modelos no lineales, MCO no es aplicable.
- Máxima verosimilitud es el método no lineal que se utiliza comunmente.
- Una ventaja es que la heterocedasticidad es tomada en cuenta automáticamente.
- Dada una muestra alaeatoria de tamaño n y una distribución $G()$, la densidad de cada y_i es:

$$f(y_i|x_i, \beta) = [G(x_i'\beta)]^{y_i}[1 - G(x_i'\beta)]^{1-y_i}$$

- Aplicando logaritmos:

$$l_i(\beta) = y_i \ln[G(x_i'\beta)] + (1 - y_i) \ln[1 - G(x_i'\beta)]$$

- Para toda la muestra, la función de verosimilitud es:

$$lnL(\beta) = \sum_{i=1}^{n} li(\beta)$$

- El estimador de β que maximiza a $L(\beta)$ es el MV, que será el estimador logit si $G()$ es la fda de una logística y el estimador probit si es una fda normal.
- El estimador es consistente, asintóticamente normal y asintóticamente eficiente.
- Para contrastar hipótesis se utilizan los estidísticos de Wald, multiplicador de lagrange y el estadístico de razón de verosimilitud.

Interpretación de las estimaciones

- Los coeficientes dan los signos de los efectos parciales de cada x_j sobre la probabilidad de respuesta.
- Existen distintas medidas de bondad de ajuste.
  - Porcentaje correctamente predicho: porcentaje de veces que $\hat{y} = y$
  - Pseudo R cuadrado: $pseudo - R^2 = 1 - \frac{1}{1+2(logL_1 - logL_0)/n}$
  - R^2 *de* $McFadden = 1 - logL_1/logL_0$
- Donde $logL_1$ es el valor de máxima verosimilitud del modelo y $logL_0$ el valor de máxima verosimilitud de un modelo solo con constante, con:

$$logL_0 \leq logL_1 < 0$$

- El efecto de x_j sobre $p(y = 1|x)$ es:

$$\frac{\partial \hat{p}(X)}{\partial x_j} = g(X\hat{\beta})\hat{\beta}_j$$

- Donde $g()$ es un factor de escala que debe ser calculado. Dos factores son:

1. Reemplazo de las variables explicativas por sus promedio muestrales

$$g(X\hat{\beta}) = g(\bar{X}\hat{\beta})$$

- Cuando se multiplica por β_j se obtiene el llamado **efecto parcial en el promedio**.
- Deficiencias
  a) Se dificulta la interpretación cuando existen variables independientes discretas.
  b) También se dificulta cuando aparecen variables en forma funcional no lineal.

2. Promediar los efectos parciales individuales a traves de la muestra.

- Cuando se multiplica por β_j se obtiene el **efecto parcial promedio** (EPP)

$$n^{-1} \sum_{i=1}^{n} g(X\hat{\beta})$$

Comparación de modelos logit, probit y MPL

- En el modelo probit se tiene que: $g(0) \approx 0.4$
- En el modelo logit se tiene que : $g(0) \approx 0.25$
- Para comparar el probit y el logit: $1.6\beta_{probit} = \beta_{logit}$
- Para comparaciones con el MPL:

$$\frac{\beta_{probit}}{2.5} = \beta_{MPL}$$

$$\frac{\beta_{logit}}{2.5} = \beta_{MPL}$$

Ejemplo: Determinantes de la participación femenina en la fuerza de trabajo

- Estime el siguiente modelo de participación en la fuerza de trabajo.

$$inlf_i = \beta_0 + \beta_1\ educ + \beta_2\ inglab + \beta_3\ kidslt6 + \beta_4 kidsge6 + \epsilon$$

```r
datos <- read.csv("mroz.csv")
attach(datos)


Y<- inlf
X = cbind(educ, nwifeinc,kidsge6,kidslt6)

stargazer(cbind(inlf,X),type="text",summary=TRUE,digits = 1,font.size = "footnotesize",header = FALSE,title="Estadísticas descri
```

```
##
## Estadísticas descriptivas
## ====================================
## Statistic  N  Mean St. Dev.  Min  Max
## ------------------------------------
## inlf      753 0.6    0.5      0    1
## educ      753 12.6   2.3      5    17
## nwifeinc  753 20.1   11.6   -0.03 96.0
## kidsge6   753 1.4    1.3      0    8
## kidslt6   753 0.2    0.5      0    3
## ------------------------------------
```

```r
#Modelo de Probabilidad Lineal
educ <- datos$educ

mpl <- lm(inlf ~ educ+ nwifeinc+kidsge6+kidslt6)

#Modelo Probit

mprobit <- glm(inlf~educ+ nwifeinc+kidsge6+kidslt6,family=binomial(link="probit"))

#Modelo Logit
mlogit <- glm(inlf~educ+ nwifeinc+kidsge6+kidslt6,family=binomial(link="logit"))

#Resumiendo Resultados

stargazer(mpl,mprobit,mlogit,type="text",digits = 2,font.size = "scriptsize", header=FALSE,no.space = TRUE)
```

```
##
## ===============================================================
##                               Dependent variable:
##                     ------------------------------------------
##                                        inlf
##                               OLS            probit  logistic
##                               (1)             (2)      (3)
## ---------------------------------------------------------------
## educ                        0.06***         0.17***  0.28***
##                              (0.01)          (0.02)   (0.04)
## nwifeinc                    -0.01***        -0.02*** -0.04***
##                             (0.002)         (0.005)   (0.01)
## kidsge6                      0.01            0.04     0.07
##                              (0.01)          (0.04)   (0.06)
## kidslt6                     -0.23***        -0.66*** -1.09***
##                              (0.03)          (0.10)   (0.17)
## Constant                     0.05           -1.32*** -2.17***
##                              (0.10)          (0.28)   (0.47)
## ---------------------------------------------------------------
## Observations                  753             753      753
## R2                            0.12
## Adjusted R2                   0.12
## Log Likelihood                               -464.82  -464.92
## Akaike Inf. Crit.                             939.64   939.85
## Residual Std. Error    0.47 (df = 748)
## F Statistic         26.09*** (df = 4; 748)
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```r
## Probabilidades predichas
```

```r
pred_mpl = predict(mpl)
pred_mlogit = predict(mlogit,type="response")
pred_mprobit = predict(mprobit,type="response")

summary(pred_mpl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.4372  0.4712  0.5953  0.5684  0.6685  1.0038
```

```r
summary(pred_mprobit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00258 0.46021 0.60272 0.56929 0.68301 0.92827
```

```r
summary(pred_mlogit)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.009788 0.455700 0.603292 0.568393 0.684940 0.917434
```

```r
## Porcentaje correctamente predicho

table(true = inlf, pred = round(fitted(mpl)))
```

```
##     pred
## true   0   1
##    0 142 183
##    1  78 350
```

```r
table(true = inlf, pred = round(fitted(mprobit)))
```

```
##     pred
## true   0   1
##    0 143 182
##    1  83 345
```

```r
table(true = inlf, pred = round(fitted(mlogit)))
```

```
##     pred
## true   0   1
##    0 145 180
##    1  82 346
```

```r
## Porcentaje correctamente predicho

table(true = inlf, pred = round(fitted(mpl)))
```

```
##     pred
## true   0   1
##    0 142 183
##    1  78 350
```

```r
table(true = inlf, pred = round(fitted(mprobit)))
```

```
##     pred
## true   0   1
##    0 143 182
##    1  83 345
```

```r
table(true = inlf, pred = round(fitted(mlogit)))
```

```
##     pred
## true   0   1
##    0 145 180
##    1  82 346
```

```r
## Pseudo R cuadrado de MacFadden
mprobit0 = update(mprobit, formula= Y ~ 1)
McFadden_probit = 1-as.vector(logLik(mprobit)/logLik(mprobit0))
McFadden_probit
```

```
## [1] 0.09721618
```

```r
mlogit0 = update(mlogit, formula= Y ~ 1)
McFadden_logit = 1-as.vector(logLik(mlogit)/logLik(mlogit0))
McFadden_logit
```

```
## [1] 0.09701092
```

```r
## Efectos marginales

#MPL
coef(mpl)
```

```
##  (Intercept)         educ     nwifeinc      kidsge6      kidslt6
##  0.048494503  0.057913785 -0.007811034  0.014237694 -0.225978819
```

```r
#Probit
gprobit = mean(dnorm(predict(mprobit, type = "link")))
gprobit * coef(mprobit)
```

```
##  (Intercept)         educ     nwifeinc      kidsge6      kidslt6
## -0.463927996  0.059380224 -0.008215485  0.014330290 -0.234097444
```

```r
#Logit
glogit = mean(dnorm(predict(mlogit, type = "link")))
glogit * coef(mlogit)
```

```
## (Intercept)        educ    nwifeinc     kidsge6     kidslt6
## -0.65306737  0.08354908 -0.01170370  0.02023043 -0.32700201
```

```r
## Evaluación de Ho: Hijos no tienen un efecto significativo sobre la participación

library(aod)
l =cbind(0,0,0,1,0)
wald.test(b=coef(mprobit), Sigma=vcov(mprobit), L=l)
```

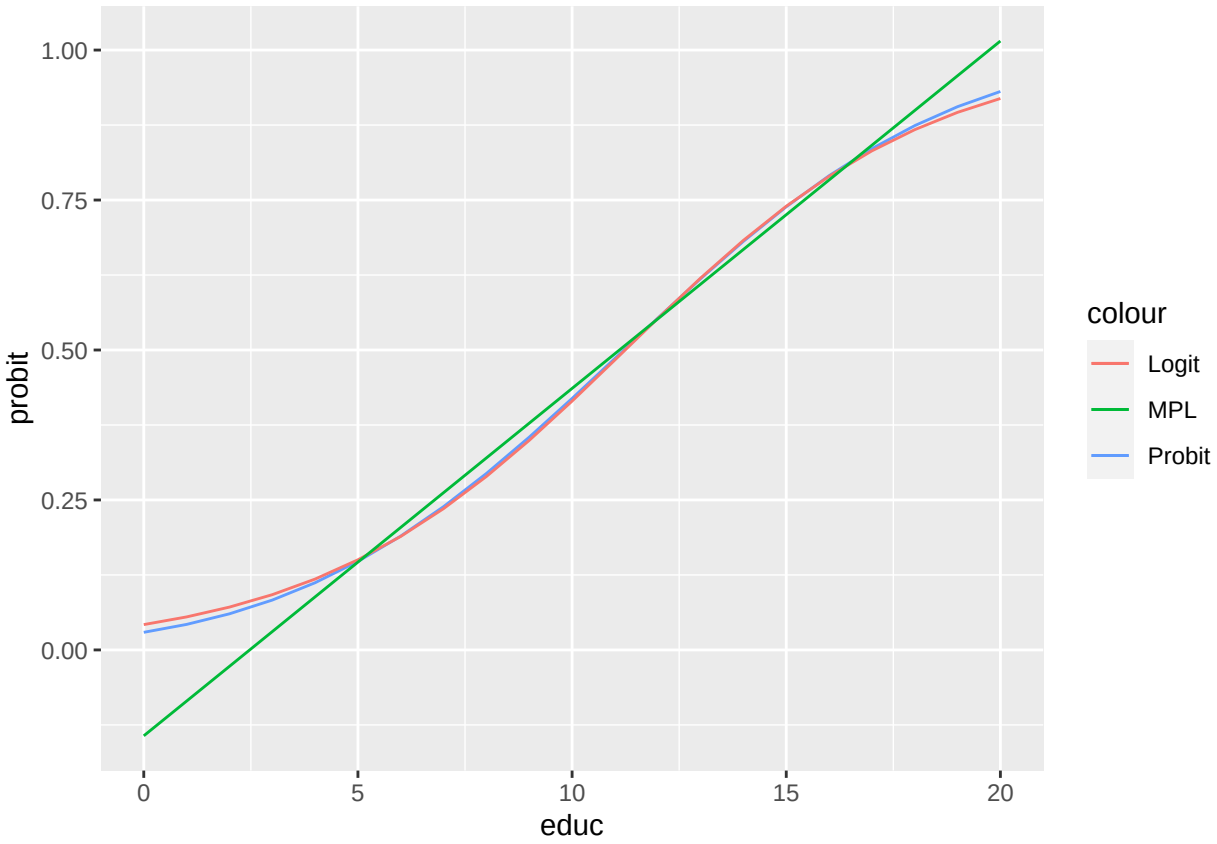
```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 1.2, df = 1, P(> X2) = 0.27
```

```r
## Simulando el impacto de cambios en la educacion sobre la participacion

nuevos_datos = data.frame(educ = seq(0,20,1),nwifeinc=rep(mean(nwifeinc),each=21)
                          ,kidsge6=rep(mean(kidsge6),each=21),kidslt6=rep(mean(kidslt6),each=21))

nuevos_datos[,c("probit")] = predict(mprobit, nuevos_datos,type="response")
nuevos_datos[,c("logit")] = predict(mlogit, nuevos_datos,type="response")
nuevos_datos[,c("mpl")] = predict(mpl, nuevos_datos,type="response")

library(ggplot2)
ggplot(data=nuevos_datos, aes(educ)) + geom_line(aes(y=probit,colour="Probit")) +
  geom_line(aes(y=logit,colour="Logit"))+geom_line(aes(y=mpl,colour="MPL"))
```



Apendice

Repaso de Estadística

Este apéndice revisa un poco de la teoría estadística y de distribución que es utilizada en estos apuntes. Más detalles pueden ser encontrados en. . .

Variables aleatorias discretas

Una **variable aleatoria** es una variable que puede tomar diferentes resultados dependiendo del "estado de la naturaleza". Por ejemplo, el resultado de lanzar una vez un dado es aleatorio, con posibles resultados 1,2,3,4,5 y 6. Sea denotada una variable aleatoria arbitraria Y. Si y denota el resultado del experimento del dado (y el dado es justo y lanzado aleatoriamente), la **probabilidad** de cada resultado es $1/6$. Esto se puede denotar como

$$P\{Y = y\} = 1/6$$

para $y = 1, 2, ..., 6$

La función que vincula los posibles resultados (en este caso $y = 1, 2, ..., 6$) a las probabilidades correspondientes es **la función de masa de probabilidad**, o más generalmente, la función de distribución de probabilidad. Esta puede ser denotada como:

$$f(y) = P\{Y = y\}$$

Note que $f(y)$ no es una función de la variable aleatoria Y, sino de todos sus posibles resultados. La función $f(y)$ tiene la propiedad que, si sumamos sobre los posibles resultados, el resultados es uno. Esto es,

$$\sum_j f(y_j) = 1$$

El **valor esperado** de una variable aleatoria discreta es el promedio ponderado de todos los posibles resultados, donde los pesos corresponden a la probabilidad de un evento particular. Se denota como:

$$E\{Y\} = \sum_j y_j f(y_j)$$

Note que $E\{Y\}$ no necesariamente corresponde a uno de los posibles resultados. En el experimento del dato, por ejemplo, el valor esperado es 3.5.

Una distribuciÃ³n es **degenerada** si está concentrada en solo un punto, esto es, si $P\{Y = y\} = 1$ para un valor particular de y y cero para el resto de los otros valores.

Variables aleatorias continuas

Una **variable aleatoria continua** puede tomar un número infinito de diferentes resultados. Por ejemplo, cualquier valor en el intervalo $[0, 1]$. En este caso, cada resultado individual tiene una probabilidad de cero. En lugar de una función de masa de probabilidad, se define la **función de densidad de probabilidad** $f(y) \geq 0$ como

$$P\{a \leq Y \leq b\} = \int_a^a f(y)dy$$

En un gráfico, $P\{a \leq Y \leq b\}$ es el área bajo la función $f(y)$ entre los puntos a y b. Tomando la integral de $f(y)$ sobre todos los posibles resultados da:

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

Si Y toma valores solamente en un determinado rango, se asume implícitamente que $f(y) = 0$ en cuaquier lugar fuera de este rango.

Se puede definir la **función de densidad acumulada** (CDF) como

$$F(y) = P\{Y \leq y\} = \int_{-\infty}^{y} f(t)dt$$

Es fÃ¡cil demostrar que $P\{a \leq Y \leq b\} = F(b) - F(a)$.

El **valor esperado** o **media** de una variable aleatoria continua, denotado usualmente como μ, es definido como

$$\mu = E\{Y\} = \int_{-\infty}^{\infty} yf(y)dy$$

Otra medida de ubicaciÃ³n es la **mediana**, que es el valor m para el que se tiene

$$P\{Y \leq m\} \geq 1/2$$

y

$$P\{Y \geq m\} \leq 1/2$$

De tal manera 50% de las observaciones están debajo de la mediana y 50% por encima. La **moda** es simplemente el valor para el cual $f(y)$ es máxima.

Una distribuciÃ³n es **simétrica** alrededor de la media si $f(\mu - y) = f(\mu + y)$. En este caso la media y la mediana de la distribuciÃ³n son idénticas.

Expectativas y momentos

Si Y y X son variables aleatorias y a y b son constantes, se tiene que

$$E\{aY + bX\} = aE\{Y\} + b\{X\}$$

lo que muestra que el valor esperado es un operador lineal. Este resultado no se sostiene necesariamiento si se consideran transformaciones no lineales de una variable aleatoria. Para un función no lineal g, no se tiene en general que $E\{g(Y)\} = g(E\{Y\})$. Si g es cÃ³ncava ($g''(Y) < 0$), la **desigualdad de Jensen**.

Por ejemplo, $E\{log(Y)\} \leq log(E\{Y\})\}$. La implicación de este es que no se puede determinar el valor esperado de una función de Y del valro esperado de Y solamente. Por definicion se cumple:.

$$E\{g(Y)\} = \int_{-\infty}^{\infty} g(y)f(y)dy$$

La **varianza** de una variable aleatoria, denotadata por σ^2, es una medida de la dispersión de la distribución. Es definida como:

$$\sigma^2 = V\{Y\} = E\{(Y-\mu)^2\}$$

igual al valor esperado de las desviacion respecto de la media al cuadrado. Es algunas veces llamada **segundo momento central**. Un resultado Ãºtil es:

$$E\{(Y-\mu)^2\} = E\{Y^2\} - 2E\{Y\}\mu + \mu^2 = E\{Y^2\} - \mu^2$$

donde $E\{Y^2\}$ es el segundo momento. Si Y tiene una distribución discreta, su varianza es determinada por:

$$V\{Y\} = \sum_j (y_j - \mu)^2 f(y_j)$$

donde j indexa los diferentes resultados posibles. Para una distribución contínua se tiene:

$$V\{Y\} = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy$$

Usando estas definiciones, es favcil verificar que:

$$V\{aY+b\} = a^2 V\{Y\}$$

donde a y b son constantes arbitrarias. Con frecuencia se utiliza la **desviación estándar** de una variable aleatoria, denotada por σ, y definida como la raíz cuadrada de la varianza. La desviación estándar es expresada en las mismas unidades que Y.

En la mayoría de los casos la distribución de una variable aleatoria no es completamente descrita por su media y su varianza, y se puede definir el momento k central como:

$$E\{(Y-\mu)^2\}, \quad k = 1, 2, 3, \cdots$$

En particualr, el tercer momento central es una medida de asimetría de la distribución alrededor de su media, mientras que el cuarto momento que mide el apuntamiento de la distribución. Típicamente, la **asimetría o *skewness*** es definido como $S \equiv E\{(Y-\mu)^3\}/\sigma^3$, mientras que la **kurtosis** es definida como $K \equiv E\{(Y-\mu)^4\}/\sigma^4$. La curtosis de una distribución normal es 3, tal que $K-3$ es referida como **exceso de curtosis**. Una distribución con exceso de curtosis positiva es llamada leptocúrtica.

Distribuciones Multivariables

La **función de distribución conjunta** de dos variables aleatorias Y y X, denotada por $f(y,x)$ es definido como:

$$P\{a_1 < Y < b_1, a_2 < X < b_2\} = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(y,x)dydx.$$

Si Y y X son **independientes**, se sostiene que $f(y,x) = f(y)f(x)$, tal que:

$$P\{a_1 < Y < b_1, a_2 < X < b_2\} = P\{a_1 < Y < b_1\}P\{a_2 < X < b_2\}$$

En general, la **distribución marginal** de Y es caracterizado por la función de densidad

$$f(y) = \int_{-\infty}^{\infty} f(y,x)dx$$

Esto implica que el valor esperado de Y es dado por:

$$E\{Y\} = \int_{-\infty}^{\infty} yf(y)dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(y,x)dxdy.$$

La **covarianza** entre Y y X es una medida de la dependencia lineal entre dos variables. Es definido como:

$$\sigma_{xy} = cov\{Y, X\} = E\{(Y - \mu_y)(X - \mu_x)\}.$$

donde $\mu_y = E\{Y\}$ y $\mu_x = E\{X\}$. El **coeficiente de correlación** es dada por la covarianza estandarizada por dos desviaciones estándar, esto es,

$$\rho_{yx} = \frac{cov\{Y, X\}}{\sqrt{V\{Y\}V\{X\}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

El coeficiente de correlación está siempre entre -1 y 1 y no es afectado por el escalamiento de las variables. El cuadrado del coeficiente de correlación entre 0 y 1 y describe la proporción de varianza en común entre Y y X. Puede ser multiplicado por 100 y expresado como porcentaje. Si $cov\{Y, X\} = 0$, Y y X se dice que **no están correlacionada**.

Cuando a, b, c, d son constantes, se sostiene que:

$$cov\{aY + b, cX + d\} = ac\ cov\{Y, X\}$$

Más aun,

$$cov\{aY + bX, X\} = a\ cov\{Y, X\} + b\ cov\{X, X\} = a\ cov\{Y, X\} + bV\{X\}$$

También se tiene que dos variables Y y X son perfectamente correlacionado ($\rho_{xy} = 1$) si $Y = aX$ para valores diferente de cero de a. Si Y y X estan correlacinados, la varianza de una función linela de Y y X depende de su covarianza. En particular,

$$V\{aY + bX\} = a^2V\{Y\} + b^2V\{X\} + 2ab\ cov\{Y, X\}$$

Si se considera un vector K-dimensional de vector de variables aleatorias, $\vec{Y} = (Y_1, \cdots, Y_K)'$, se puede definir su vector de expectativas:

$$E\{\vec{Y}\} = \begin{pmatrix} E\{Y_1\} \\ \vdots \\ E\{Y_K\} \end{pmatrix}$$

y su matriz varianza covarianza (o simplemente **matriz de covarianza**) como:

$$V\{\vec{Y}\} = \begin{pmatrix} V\{Y_1\} & \cdots & cov\{Y_1, Y_K\} \\ \vdots & \ddots & \vdots \\ cov\{Y_K, Y_1\} & \cdots & V\{Y_K\} \end{pmatrix}$$

Note que esta matriz es simétrica. Si consideramos una o más combinaciones lineales de los elementos en $\vec{Y}$, es decir $R\vec{Y}$, dibde R es de dimensión $J \times K$ se tiene que:

$$V\{R\vec{Y}\} = RV\{\vec{Y}\}R'$$

Distribuciones Condicionales

Una distribución condicional describe la distribución de una variable, diga Y, dada la realización de otra variable X. Por ejemplo, si se lanza un dado dos veces, X puede denotar la realización del primer dado y Y puede denotar el total de dos dados. Entonces se puede estar interesado en la distribución de Y condicional a la realización del primer dado. Por ejemplo, cuál es la probabilidad de lanzar 7 en total si la realización del primer dado es 3? O una realización de 3 o menos? La distribución condicional está implícita por la distribución conjunta de dos variables. Se define,

$$f(y|X = x) = f(y|x) = \frac{f(y, x)}{f(x)}$$

Si Y y X son independiente, inmediatamente sigue que $f(y|x) = f(y)$. De la definición anterior sigue que:

$$f(y, x) = f(y|x)f(x),$$

que dice que la distribución conjunta de dos variables puede ser descompuesta en el producto de una distribución condicional y una distribución marginal. De manera similar, se puede escribir:

$$f(y,x) = f(x|y)f(y).$$

La **esperanza condicional** de Y dado $X = x$ es el valor esperado de Y de la distribución condicional. Esto es,

$$E\{Y|X = x\} = E\{Y|x\} = \int yf(y|x)dy$$

La expectativa condicional es una función de x, a menos que Y y X son independientes. De manera simular, se puede definir la varianza condicional como:

$$V\{Y|x\} = \int (y - E\{Y|x\})^2 f(y|x)dy$$

que puede ser escrita como:

$$V\{Y|x\} = E\{Y^2|x\} - (E\{Y|x\})^2.$$

Se tiene que:

$$V\{Y\} = E_x\{V\{Y|X\}\} + V_x\{E\{Y|X\}\},$$

donde E_x y V_x denotan el valor esperado y varianza, respectivamente, obtenida de la distribución marginal de X. Los términos $V\{Y|X\}$ y $E\{Y|X\}$ son funciones de la variable aleatoria X y en consecuencia son variables aleatorias en sí mismas.

Considerese la relación entre dos variables aleatorias Y y X, donde $E\{Y\} = 0$. Entonces, se tiene que Y y X están **no correlacionada** si:

$$E\{YX\} = cov\{Y, X\} = 0$$

Si Y es **independiente en media condicional** de X, significa que:

$$E\{Y|X\} = E\{Y\} = 0.$$

Este resultado es más fuerte que correlación cero porque $E\{Y|X\} = 0$ implica que $E\{Yg(X)\} = 0$ para cualquier función g. Si Y y X son **independientes**, este es también fuerte e implica que:

$$E\{g_1(Y)g_2(Y)\} = E\{g_1(Y)|E\{g_2(Y)\},$$

para funciones arbitrarias g_1 y g_2. Es facilmente verificable que esto implica independencia en media condiciona y correlación cero. Note que $E\{Y|X\} = 0$ no necesariamente implica que $E\{X|Y\} = 0$.

La distribución normal

En econometría, la **distribución normal** juega en un rol central. La función de densidad para una distribución normal con media μ y varianza σ^2 es dada por:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\}$$

donde se puede escribir como $Y \sim N(\mu, \sigma^2)$. Es fácil verificar que la distribución normal es simétrica.

Una distribución normal estándar es obtenida con $\mu = 0$ y $\sigma = 1$. Note que la variable estandarizada $(Y - \mu)/\sigma$ es $N(0, 1)$ si $Y \sim N(\mu, \sigma^2)$. La densidad de una distribución normal, típicamente denotada por ϕ, está dada por:

$$\phi(y) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}y^2\}$$

Una propiedad útil de la distribución normal es que una función lineal de una variable normal es también normal. Esto es, si $Y \sim N(\mu, \sigma^2)$, entonces:

$$aY + n \sim N(a\mu + b, a^2\sigma^2)$$

La función de densidad acumulada de la distribución normal no tienen forma cerrada. Se tienen que:

$$P\{Y \leq y\} = P\{\frac{Y-\mu}{\sigma} \leq \frac{y-\mu}{\sigma}\} = \Phi\left(\frac{y-\mu}{\sigma}\right) = \int_{\infty}^{(y-\mu)/\sigma} \phi(t)dt$$

donde Φ denota la CDF de una distribución normal estándar. Note que $\Phi(y) = 1 - \Phi(-y)$ debido a la simétria.

La simetría también implica que el tercer momento central de una distribución normal es cero. Puede mostrarse que el cuarto momento de una distribución normal está dada por:

$$E\{(Y-\mu)^4\} = 3\sigma^2.$$

Típicamente estas propiedades del tercer y cuarto momento son explotadas en los constrastes de normalidad.

Si (Y, X) tiene una **distribución normal bivariada** con vector de medias $\mu = (\mu_y, \mu_x)$ y matriz de covarianza:

$$\sum = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}$$

denotado por $(Y, X)' \sim N(\mu, \sum)$, la distribución conjuta está dada por:

$$f(y, x) = f(y|x)f(x)$$

donde ambas la **densidad condicional** de Y dado X y la **densidad marginal** de X son normales. La densidad condicional está dada por:

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma_{y|x}^2}} \exp\{-\frac{1}{2}\frac{(y-\mu_{y|x})^2}{\sigma_{y|x}^2}\}$$

donde $\mu_{y|x}$ es la **expectativa condicional** de Y dado X, dada por:

$$\mu_{y|x} = \mu_y + (\sigma_{yx}/\sigma_x^2)(x - \mu_x),$$

y $\sigma_{y|x}^2$ es la varianza condicional de Y dado X,

$$\sigma_{y|x}^2 = \sigma_y^2 - \frac{\sigma_{yx}}{\sigma_x^2} = \sigma_y^2(1 - \rho_{yx}^2)$$

con ρ_{yx} denotando el coeficiente de correlación entre Y y X. Estos resultados tienen implicaciones importantes. Primero, si dos (o más) variables tienen una distribución conjunta normal, todas las distribuciones marginales y condicionales son también normales. Segundo, la expectativa condicional de una variable dada las demás es una función lineal (con un término intercepto). Tercero, si $\rho_{yx} = 0$, entonces $f(y|x) = f(y)$, tal que

$$f(y, x) = f(y)f(x)$$

y Y y X son independientes. Entonces si Y y X tienen una distribución conjunta con correlación cero, entonces ellas son automáticamente independientes. Recuerde que en general independencia es más fuerte que no correlación.

Otro resultado importante es que una función lineal de variables normales es también normal, que es, si $(Y, X)' \sim N(\mu, \sum)$, entonces

$$aY + bX \sim N(a\mu_y + b\mu_x, a^2\sigma_y^2 + b^2\sigma_x^2 + 2ab\sigma_{yx})$$

Estos resultados pueden ser generalizados a una distribución normal de K variables. Si el vector K dimensional $\vec{Y}$ tiene distribución normal con media el vector μ y matriz de covarianza $\sum$, esto es,

$$\vec{Y} \sim N(\mu, \Sigma)$$

se tiene que la distribución de $R\vec{Y}$, donde R es una matriz $J \times K$, es una distribución normal J-variante, dada por:

$$R\vec{Y} \sim N\left(R\mu, R\Sigma R'\right)$$

En modelos con variables dependientes limitadas se encuentran con frecuencia formas de **truncamiento**. Si Y tienen una densidad $f(y)$, la distribución truncada por debajo en un punto c ($Y \geq c$) es dada por:

$$f(y|Y \geq c) = \frac{f(y)}{P\{Y \geq c\}} \quad si \quad y \geq c \quad , \ 0 \ lo \ contrario$$

Si Y es una variable normal estándar, la distribución truncada de $Y \geq c$ tiene media

$$E\{Y|Y \geq c\} = \lambda_1(c)$$

donde:

$$\lambda_1(c) = \frac{\phi(c)}{1 - \Phi(c)},$$

y varianza:

$$V\{Y|Y \geq c\} = 1 - \lambda_1(c)[\lambda_1(c) - c].$$

Si la distribución es truncada por arriba ($Y \leq c$), se tiene que:

$$E\{Y|Y \leq c\} = \lambda_2(c),$$

con

$$\lambda_2(c) = \frac{-\phi(c)}{\Phi(c)}$$

Si Y tiene una densidad normal con media μ y varianza σ^2, la distribución truncada $Y \geq c$ tiene media:

$$E\{Y|Y \geq c\} = \mu + \sigma\lambda_1(c^*) \geq \mu$$

donde $c^* = (c - \mu)/\sigma$, y, de manera similar,

$$E\{Y|X \geq c\} = \mu_y + \left(\frac{\sigma_{yx}}{\sigma_x^2}\right)[E\{X|X \geq c\} - \mu_x]$$

$$= \mu_y + \left(\frac{\sigma_{yx}}{\sigma_x}\right)\lambda_1(c^*)$$

Otras distribuciones relacionadas

Más allá de la distribución normal, existen otras distribuciones importantes. Primero, se define la **distribución Chi cuadrada** como sigue. Si $Y_1, ..., Y_J$ es un conjunto de variables independientes con distribución normal estándar, se tiene que:

$$\xi = \sum_{j=1}^{J} Y_j^2$$

tiene una distribución Chi cuadrada con J grados de libertad. Se denota como $\xi \sim \chi_J^2$. De manera general, si un conjunto de variables normales independientes con media μ y varianza σ^2, se sigue que:

$$\xi = \sum_{j=1}^{J} \frac{(Y_J - \mu)^2}{\sigma^2}$$

es Chi cuuadrada con J grados de libertad. Aun de forma mas general, si $\vec{Y} = (Y_1, ..., YJ)'$ es un vector de variables aleatorias que tiene un distribución normal conjunta con vector de medias μ y matriz de covarianzas (no singular), sigue que:

$$\xi = (\vec{Y} - \mu)' \sum^{-1} (\vec{Y} - \mu) \sim \chi_J^2$$

Si xi tiene una distribución Chi cuadrada con J grados de libertad, se puede mostrar que $E\{\xi\} = J$ y $V\{\xi\} = 2J$.

A continuación, considere la **distribución t** (o distribución de Student). Si X tiene una distribución normal estándar, $X \sim N(0, 1)$, y $\xi \sim \chi_J^2$, y si X y ξ son independientes, la razón sigue una distribución t con J grados de libertad. Como la

distribución normal estándar, la distribución t es simétrica y centrada en cero, pero tiene colas más "gruesas", particularmente para J pequeño. Si J se aproxima a infinito, la distribución t se aproxima a la distribución normal.

Si $\xi_1 \sim \chi^2_{J_1}$ y $\xi_2 \sim \chi^2_{J_2}$, y si ξ_1 y ξ_2 son independientes, se tiene que la razón:

$$f = \frac{\xi_1/J_1}{\xi_2/J_2}$$

tiene una **distribución F** con J_1 y J_2 grados de libertad en el numerador y el denominador respectivamente. La razón inversa:

$$\frac{\xi_2/J_2}{\xi_1/J_1}$$

también tiene una distribución F, pero con J_2 y J_1 grados de libertad respectivamente. La distribución F es por tanto la distribución de la razón de dos variables Chi cuadradas independientes, dividios por sus respectivos grados de libertad. Cuando $J_1 = 1$, ξ_1 es una variable normal cuadrada, por ejemplo $\xi_1 = X^2$, entonces:

$$t^2 = \left(\frac{X}{\sqrt{\xi/J_2}} \right)^2 = \frac{\xi_1}{\xi_2/J_2} = f \sim F^1_{J_2}$$

Entonces con un grado de libertad en el numerador, la distribución F es justamente el cuadrado de la distribución t. Si J_2 es grande, la distribución de

$$J_1 f = \frac{\xi_1}{\xi_2/J_2}$$

es bien aproximada por una distribución Chi cuadrada con J_1 grados de libertad. Para un J_2 grande el denominador es insignificante.

Finalmente, considere la **distribución log-normal**. Si $\log Y$ tiene una distribución normal con media μ y varianza σ^2, entonces $Y > 0$ tiene una distribución llamada log-normal. La densidad log-normal es usualmente utilizada para describir la distribución poblacional del ingreso (laboral) o la distribución de retornos de activos. Mientras que $E\{\log Y\} = \mu$, se cumple que:

$$E\{Y\} = \exp\{\mu + \frac{1}{2}\sigma^2\}$$

.

Repaso de Algebra Matricial

Vectores y Matrices

Para fines de facilitar y hacer más transparente a notación matemática del texto se hace uso del algebra lineal.

Terminología

En este libro un **vector** es una columna de números, denotado por:

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

La **transpuesta** de un vector, denotado por $a' = \begin{pmatrix} a_1, a_2, \cdots, a_n \end{pmatrix}$, es una fila de números, algunas veces llamado vector fila.

Una **matriz** es un arreglo rectangular de números de dimensión $n \times k$ que puede ser escrita como:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{pmatrix}$$

El primer índice del elemento a_{ij} se refiere a la fila i, y el segundo índice se refiere a la columna j. Denotando el vector de la columna j por a_j, se ve que A consiste de k vectores a_1 a a_k, y puede ser denotada como:

$$A = \begin{pmatrix} a_1 & a_2 & \cdots & a_k \end{pmatrix}$$

El símbolo $'$ denota la transpuesta de una matriz o vector,

$$A' = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \cdots & & & \\ a_{1k} & a_{2k} & \cdots & a_{nk} \end{pmatrix}$$

Las columnas de A son las filas de A, y viceversa. Una matriz es **cuadrada** si $n = k$. Una matriz A es **simétrica** si $A = A'$. Una matriz cuadrada es llamada matriz **diagonal** si $a_{ij} = 0$ para todo $i \neq j$. Note que una matriz diagonal es simétrica por construccion. La **matriz identidad** I es una matriz diagonal con todos los elementos de la diagonal igual a uno.

Manipulación de matrices

Si dos matrices o vectores tienen una misma dimensión, estos pueden ser *sumados* o *sustraidos*. Sean A y B dos matrices de dimensiónes $n \times k$ con elementos a_{ij} y b_{ij}, respectivamente. Entonces $A + B$ tiene un elemento típico $a_{ij} + b_{ij}$, mientras que $A - B$ tiene el elemento típico $a_{ij} - b_{ij}$. Sigue facilmente que $A + B = B + A$ y que $(A + B)' = A' + B'$.

Una matriz A de dimensiones $n \times k$ y una matriz B de dimensiónes $k \times m$ pueden ser multiplicadas para producir una matriz de dimensiónes $n \times m$. Considerando el caso especial donde $k = 1$, entonces $A = a'$ es un vector fila y $B = b'$ es un vector columna. Entonces, se define:

$$AB = a'b = (a_1, a_2, \cdots, a_n) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Se denomina $a'b$ como **producto interno** o producto punto de los vectores a y b. Note que $a'b = b'a$. Dos vectores son llamados ortogonales si $a'b = 0$. Para cualquier vector a, excepto el vetor nulo, se tiene que $a'a > 0$. El producto exterior de un vector a es aa', que es de dimensión $n \times n$.

Un caso especial surge para $m = 1$. En este caso A es una matriz $n \times k$ y $B = b$ es un vector de dimensión k. Entonces $c = Ab$ es tambien un vector, pero de dimensión n. Sus elementos son:

$$c_i = a_{i1} b_1 + a_{i2} b_2 + \cdots + a_{ik} b_k$$

que es el producto interno entre el vector obtenido de la fila i de A y el vector b.

Cuando $m > 1$, B es una matriz y $C = AB$ es una matriz de dimensión $n \times m$ con elementos típicos

$$c_{ij} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{ik} b_{kj}$$

siendo el producto interno entre los vectores obtenidos de la fila i de A y la columna j de la matriz B. Note que este solo hace sentido si el número de columnas de A es igual al número de filas en B.

Considere el siguiente ejemplo:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 0 \end{pmatrix}$$

y

$$B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 5 \end{pmatrix}$$

y

$$AB = ?$$

Es importante notar que $AB \neq BA$. Incluso si AB existe, BA no puede estar definida, debido a que las dimensiónes de B y A no son conformables. Si A es de dimensión $n \times k$ y B es de dimensión $k \times n$, entonces AB existe y tienen dimensión $n \times n$, mientras que BA existe con dimensión $k \times k$. En el ejemplo anterior, se tiene que

$$BA = \begin{pmatrix} 9 & 12 & 3 \\ 19 & 26 & 9 \\ 20 & 25 & 0 \end{pmatrix}$$

Para la transpuesta de un producto de dos matrices, se tiene que

$$(AB)' = B'A'$$

De aquí (y $(A')' = A$) sigue que ambas $A'A$ y AA' existen y son simétricas. Finalmente, multiplicando un escalar y una matriz es lo mismo que multiplicar cada elemento en la matriz por este escalar. Esto es, para un escalar c, cA tiene como elemento típico ca_{ij}

Propiedades de las matrices y vectores

Si se considera un número de vectores a_1 a a_k, se puede tener una **combinación lineal** de estos vectores. Con pesos los escalares $c_1, \cdots, c_k$ este produce el vector $c_1a_1 + c_2a_2 + \cdots + c_ka_k$, que puede ser escrito como Ac, donde, como antes $A = [a_1, \cdots, a_2]$ y $c = (c_1, \cdots, c_k)'$.

Un conjunto de vectores es **linealmente dependiente** si cualquier de los vectores pueden ser escritos como una combinación lineal de los otros. Esto es, si existen valores para $c_1, \cdots, c_k$, no todos cero, tal que $c_1a_1 + c_2a_2 + \cdots + c_ka_k = 0$ (el vector nulo). Igualmente, un conjunto de vectores es linealmente independiente si la solución a

$$c_1a_1 + c_2a_2 + \cdots + c_ka_k = 0$$

es

$$c_1 = c_2 = \cdots = c_k = 0$$

Esto es, si la unica solución de $Ac = 0$ es $c = 0$.

Si se consideran todos los vectores posibles que pueden tenerse como combinación lineal de los vectores $a_1, \cdots, a_k$, estos vectores forman un **espacio vectorial**.

Si los vectores $a_1, \cdots, a_k$ son linealmente dependientes, se puede reducir el número de vectores sin cambiar el espacio vectorial. El número mínimo de vectores necesarios para generar un espacio vectorial es llamado **dimensión** de ese espacio. En este sentido se puede definir el **espacio columna** de una matriz como el espacio generado por sus columnas, y el **rango columna** de una matriz como la dimensión de su espacio columna. Es claro, que el rango columna nunca puede exceder el número de columnas. Una matriz es de rango columna completo si el rango columna es igual al número de columnas. El **el rango fila** de una matriz es la dimensión del espacio generado por las filas de una matriz. En general, se sostiene que el rango fila y el rango columna de una matriz son iguales, de tal manera que inambiguamente define el rango de una matriz. Note que esto no implica que una matriz que es de rango columna completo es automaticamente de rango fila completo (esto solo se cumple si la matriz es cuadrada).

Un resultado útil en el análisis de regresión para cualquier A es:

$$rango(A) = rango(A'A) = rango(AA')$$

Matrices Inversas

Una matriz B, si existe, es la inversa de la matriz A si $AB = I$ y $BA = I$. Un requerimiento necesario para esto es que A sea una matriz cuadrada y tenga rango completo. En este caso es llamada **invertible** o **no singular**. En ese sentido, se puede definir $B = A^{-1}$,

$$AA^{-1} = I$$

y

$$A^{-1}A = I$$

Note que la definición implica que $A = B^{-1}$. Entonces, se tiene que $(A^{-1})^{-1} = A$. Si A^{-1} no existe, decimos que A es **singular**. Analíticamentre, la inversa de una matriz diagonal y la inversa de una matriz 2×2 se obtienen facilmente. Por ejemplo,

$$\begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} = \begin{pmatrix} a_{11}^{-1} & 0 & 0 \\ 0 & a_{22}^{-1} & 0 \\ 0 & 0 & a_{33}^{-1} \end{pmatrix}$$

y

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

Si $a_{11}a_{22} - a_{12}a_{21} = 0$, la matriz A 2×2 es singular: sus columnas son linealmente dependienes, y por lo tanto sus filas. Se denomina a $a_{11}a_{22} - a_{12}a_{21}$ el **determinante** de esta matriz 2×2.

Suponga que se requiere resolver $Ac = d$ para un A y d datos, donde A es de dimensión $n \times n$ y ambos c y d son vectores n-dimensionales. Este es un sistema de n ecuaciones lineales con n incógnitas. Si A^{-1} existe, se puede escribir:

$$A^{-1}Ac = c = A^{-1}d$$

para obtener la solución. Si A es no invertible, el sistema de ecuaciones lineales tiene dependencias lineales. Hay dos posibilidades: más de un vector satisface $Ac = d$, de tal manera que no existe solución unica o las ecuaciones son inconsistenes, por tanto no existe solución al sistema. Si d es el vector nulo, solo la primera posibilidad es posible.

Facilmente se puede derivar que:

$$(A^{-1})' = (A')^{-1}$$

y que

$$(AB)^{-1} = B^{-1}A^{-1}$$

(asumiendo que ambas existan).

Matrices idempotentes.

Un caso especial es el de matrices simétricas e idempotentes. Una matriz P es **simétrica** si $P' = P$ e **idempotente** si $PP = P$. Una matriz simétrica idempotente P tiene la interpretacion de una **matriz de proyeccion**. Esto significa que el vector de proyeccion Px esta en el espacio columna de P, mientras que el vector residual $x - Px$ es ortogonal a cualquier vector en el espacio columna de P.

Una matriz de proyección que proyecte el espacio columna de una matriz A puede ser construido como $P = A(A'A)^{-1}A'$. Claramente, esta matriz es simétrica e idempotente.Proyectando dos veces sobre el mismo espacio, no altera el resultado, de tal manera que se tiene que $PPx = P$, que resulta directamente. El residuo de la proyeccion es $x - Px = (I - A(A'A)^{-1}A')x$ tal que $M = I - A(A'A)^{-1}A'$ es también una matriz de proyeccion con $MP = PM = 0$ y $MM = M = M'$. Entonces, los vectores Mx y Px son ortogonales.

Una matriz de proyecciones interesante, que se usará más adelante, es $Q = I - (1/n)u'$. Donde u' es una matriz de unos. Los elementos de la diagonal de esta matriz son $1 - 1/n$, y los elementos fuera de la diagonal son $-1/n$. Ahora, Qx es un vector que contiene las desviacions de x de su media. Un vector de medias es producido por la matriz de transformación $P = (1/n)u'$. Note que $PP = P$ y $QP = 0$.

La única matriz de proyección que no es singular es la matriz identidad. Las demás matrices de proyección son singulares, cada una con el rango igual a la dimensión del espacio sobre la que se hará la proyección.

Valores y vectores propios

Sea A una matriz simétrica $n \times n$. Considere el siguiente problema de encontrar combinaciónes de un vector c (otro diferente al vector nulo) y un escalar λ que satisface:

$$Ac = \lambda c$$

En general, hay n soluciónes $\lambda_1, \lambda_2, ..., \lambda_n$ llamados **valores propios** (raíces caracteristicas) de A, correspodientes a n vectores $c_1, c_2, ..., c_n$ llamados **vectores propios** (vectores característicos). Si c_1 es una solución, entonces también es cierto que kc_1 es una solución para cualquier constante k, de tal manera que los vectores propios esta definidos hasta una constante. Los vectores propios de una matriz simétrica son ortogonales, esto es, $c_i'c_j = 0$ para todo $i \neq j$.

Si un valor propio es cero, el vector correspondiente c satisface $Ac = 0$ que implica que A no es de rango completo y por lo tanto singular. Entonces, una matriz singular tiene al menos un valor propio igual a cero. En general, el rango de una matriz simétrica corresponde al número de valores propios diferentes de cero.

Una matriz simétrica es llamada **positiva definida** si todos su valores propios son positivos. Es llamada **positiva semidefinida** si todos sus valores propios son no negativos. Una matriz positiva definida es invertible. Si A es positiva definida, se tiene que para cualquier vector de x (diferente al vector nulo) que

$$x'Ax > 0$$

La razón es que cualquier vector x puede ser escrito como una combinación lineal de los vectores propios como $x = d_1c_1 + ... + d_nc_n$ para los escalares $d_1, ..., d_n$, y se puede escribir

$$x'Ax = (d_1c_1 + ... + d_nc_n)'A(d_1c_1 + ... + d_nc_n) = \lambda_1 d_1^2 c_1'c_1 + ... + \lambda_n d_n^2 c_n'c_n > 0$$

De manera similar, para una matriz A positiva semi-definida, se tiene para cualquier vector x

$$x'Ax \geq 0$$

El **determinante** de una matriz simétrica es igual al producto de sus n valores propios. El determinante de una matriz positiva definida es positivo. Una matriz simétrica es singular si el determinante es cero (eso es, si uno de los valores propios es cero).

Diferenciación

Sea x un vector columna n-dimensiónal. Si c es también un vector columna n-dimensiónal, $c'x$ es un escalar. Considerese $c'x$ como una función del vector x. Entonces, se puede considerar el vector de derivada de $c'x$ con respecto a cada uno de los elementos en x, esto es,

$$\frac{\partial c'x}{\partial x} = c$$

Este es un vector columna con n derivadas, donde el elemento típico es c_i. De manera más general, para una función vectorial Ax (donde A es una matriz) se tiene:

$$\frac{\partial Ax}{\partial x} = A'$$

El elemento en la columna i, fila j de la matriz es la derivada del elemento j en la función Ax respecto a x_j.

Más aun,

$$\frac{\partial x'Ax}{\partial x} = 2Ax$$

para una matriz A simétrica. Si A no es simétrica, se tiene que

$$\frac{\partial x'Ax}{\partial x} = (A + A')x$$

Todos estos resultados resultan de coleccionar los resultados de una diferenciación elemento por elemento.

Algunas manipulaciones es de mínimos cuadrados:

Sea $x_i = (x_{i1}, x_{i2}, ..., x_{iK})'$ con $x_{i1} \equiv 1$ y $\beta = (\beta_1, \beta_2, ..., \beta_K)'$. Entonces,

$$x_i'\beta = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

La matriz

$$\sum_{i=1}^{N} x_i x_i' = \sum_{i=1}^{N} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix} (x_{i1}, x_{i,2}, \cdots, x_{iK})$$

$$= \begin{pmatrix} \sum_{i=1}^N x_{i1}^2 & \sum_{i=1}^N x_{i2}x_{i1} & \cdots & \sum_{i=1}^N x_{iK}x_{i1} \\ \vdots & \sum_{i=1}^N x_{i2}^2 & & \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^N x_{i1}x_{iK}) & & \cdots & \sum_{i=1}^N x_{iK}^2 \end{pmatrix}$$

es una matriz $K \times K$ simétrica que contiene la suma de cuadrados y los productos cruzados. El vector:

$$\sum_{i=1}^{N} x_i y_i = \begin{pmatrix} \sum_{i=1}^{N} x_{i1} y_i \\ \sum_{i=1}^{N} x_{i2} y_i \\ \vdots \\ \sum_{i=1}^{N} x_{iK} y_i \end{pmatrix}$$

tiene tamaño K, de tal manera que el sistema

$$\left(\sum_{i=1}^{N} x_i x_i' \right) b = \sum_{i=1}^{N} x_i y_i$$

es un sistema de K ecuaciones con K incógnitas (en b). Si $\sum_{i=1}^{N} x_i x_i'$ es invertible, una solución unica existe. La invertibilidad requiere que $\sum_{i=1}^{N} x_i x_i'$ sea de rango completo. Si no es de rango completo, existe un vector diferente de cero y de dimensión K, tal que $x_i' c = 0$ para cada i, y existe una relación lineal entre las columnas/filas de la matriz $\sum_{i=1}^{N} x_i x_i'$.

En notación matricial, la matriz $N \times K$ X es definida como:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{iK} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}$$

y $y = (y_1, y_2, \cdots, y_N)'$. De ahi se puede verificar que

$$X'X = \sum_{i=1}^{N} x_i x_i'$$

y

$$X'y = \sum_{i=1}^{N} x_i y_i$$